university of groningen

campus fryslân

Minimal Acoustic Markers for Age Prediction in Human Voice: A Machine Learning Approach

Hiva Naazeri

S6028497

Supervisor: Prof. Dr. Matt Coler

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Minimal Acoustic Markers for Age Prediction in Human Voice: A Machine Learning Approach

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
Assoc. Prof. Dr. Matt Coler (Voice Technology, University of Groningen)

Hiva Naazeri (S6028497)

June 11, 2025

Acknowledgements

First and foremost, I want to express my deepest gratitude to my supervisor, Assoc. Prof. Dr. Matt Coler. If you're reading this thesis, know that it wouldn't exist without his guidance, support, and encouragement. From the very beginning to the final draft, he was a constant source of inspiration, insight, and motivation. Quite simply—he did everything. I am incredibly thankful for his dedication and belief in this work.

I'm also sincerely grateful to Campus Fryslân and the University of Groningen. If you've ever studied or visited, you'll know how much the environment fosters curiosity, creativity, and critical thinking. The opportunities and support here have shaped every part of this journey.

To all the professors of the Master of Voice Technology program, thank you for sharing your expertise, enthusiasm, and support throughout this journey. Your dedication to the field and to our learning has left a lasting impression, and I feel fortunate to have been part of such a passionate academic community.

To my aunt, Dr. Sepideh Yousefzadeh, thank you for your unwavering support and encouragement. Your guidance and kindness have been a great source of strength throughout my studies.

I would also like to thank Mohammadhossein Narang for his valuable support and encouragement during this journey.

To my parents—thank you for your unwavering love and support, even from afar. Your encouragement and belief in me, especially during the more challenging moments, have meant everything. I carry your strength with me in every step I take.

And finally, thank you for taking the time to read this thesis. I hope it brings you something useful, interesting, or inspiring—just as the people behind it brought all those things to me.

Abstract

Aging affects the human voice in systematic and measurable ways due to physiological changes in the vocal tract, respiratory system, and laryngeal structures. This thesis investigates the extent to which vocal characteristics can be used to predict a speaker's age group using a minimal, interpretable set of biologically motivated acoustic features. Leveraging a curated subset of the Mozilla Common Voice dataset, we extracted features such as fundamental frequency (F0), formant frequencies, jitter, shimmer, spectral tilt, speech rate, and mel-frequency cepstral coefficients (MFCCs) to train machine learning models for age group classification.

We developed a reproducible audio processing and feature extraction pipeline using open-source tools and evaluated several models, with Random Forests demonstrating the best performance, achieving up to 62% accuracy across five broad age groups. Feature importance analysis revealed that vocal perturbation measures (jitter and shimmer), spectral features, and speech rate were among the most informative for predicting speaker age. Despite limited accuracy in underrepresented age groups (e.g., 50s and 60s), the results suggest that interpretable acoustic biomarkers capture meaningful age-related vocal changes.

This work provides a baseline for age prediction from voice with practical implications in human-computer interaction, speaker profiling, and health monitoring. Limitations include class imbalance, reliance on self-reported age labels, and language-specific data. Future research should explore data augmentation, continuous age prediction via regression, expanded feature sets, and cross-linguistic generalizability. Clinical extensions include using vocal biomarkers for early detection of age-related diseases and neurodegenerative disorders, offering a promising, non-invasive diagnostic avenue.

Keywords: Voice Aging, Acoustic Biomarkers, Age Prediction, Speech Processing, Machine Learning, Jitter, Shimmer, Spectral Features, Random Forest, Vocal Health Monitoring

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 (from any source: printed, internet or other), this has been carefully acknowledged and referenced. During the preparation of this thesis, I used ChatGpt-3.5, ChatGpt-o4-mini and Gemini (2.5 flash) for the following purposes:
sentence restructuring in chapters 1, 2, 3, 4, 5, 6, and 7, generating alternative explanations for technical concepts in chapters 3, 4, 5, creating initial code documentation templates, summarizing background literature for preliminary review. All content was subsequently reviewed, verified, and substantially modified by me.

Hiva Naazeri / 11 June 2025

Contents

1 Introduction

In an era marked by rapid population aging, the importance of understanding physiological and behavioral markers of aging has become more than a scientific curiosity—it is a societal necessity. According to the World Health Organization (2021), by 2050 the number of individuals aged 60 years and older will reach 2.1 billion, representing over 20% of the global population. This demographic transformation poses significant challenges to healthcare systems, labor markets, and social infrastructure. In parallel, it presents opportunities for early screening tools and preventive strategies that can support aging individuals in maintaining functional independence and quality of life. A growing body of interdisciplinary research has explored various biomarkers of aging, ranging from cellular and molecular indicators to observable phenotypic shifts. Yet, the human voice—an accessible, non-invasive, and information-rich channel—remains underutilized in the scientific toolkit for age-related monitoring.

Speech production involves intricate coordination among the respiratory, laryngeal, and articulatory systems, all of which undergo biological changes with age. These anatomical and physiological shifts manifest acoustically, altering the voice in ways that are both perceptible to listeners and measurable via signal processing techniques. Research in speech science and phonetics has documented characteristic changes in vocal parameters over the adult lifespan. Common age-related acoustic changes include a decrease in fundamental frequency (F0) in females and an increase in males due to hormonal and anatomical factors (Linville, 2002), increased perturbation measures such as jitter and shimmer due to declining vocal fold control (Keerthiga and Shetty,2023), a shift in spectral tilt due to weakening respiratory support (Goy, Fernandes, Pichora-Fuller, and van Lieshout,2013), and a general slowing of speech rate (Torre and Barlow, 2009). These changes are not only physiologically grounded but also quantifiable, making them prime candidates for computational modeling.

Despite these well-documented shifts, machine learning models for voice-based age prediction have largely emphasized performance over interpretability, often relying on high-dimensional feature spaces with limited biological relevance. For instance, deep learning models frequently utilize large sets of mel-frequency cepstral coefficients (MFCCs), spectrograms, or embeddings from pretrained networks (e.g., wav2vec or x-vectors), which may achieve high classification accuracy but obscure the contribution of individual acoustic dimensions (Eyben, Wöllmer, and Schuller, 2015;Li, Peng, Wang, Li, and Wu, 2022). Such approaches also tend to be data-hungry and computationally expensive, limiting their applicability in low-resource or real-time contexts. Furthermore, most models are trained on datasets with limited demographic diversity or are biased toward younger speakers, reducing generalizability across the older adult population (Schuller, Steidl, Batliner, et al., 2013).

There is a growing need to develop lightweight, interpretable models for age prediction from speech that are not only accurate but also grounded in vocal physiology. This study proposes a novel approach by focusing on a minimal set of biologically motivated acoustic features. Specifically, we examine six acoustic parameters—fundamental frequency (F0), formant dispersion (F1–F3), jitter, shimmer, spectral tilt, and speech rate—that have been empirically linked to anatomical changes in the larynx and vocal tract. Each feature corresponds to a known physiological mechanism. For example, F0 reflects vocal fold tension and length, which tend to change with age due to hormonal and muscular shifts; jitter and shimmer capture micro-instabilities in vocal fold vibration that increase

with neuromuscular degradation; spectral tilt reflects the balance between high- and low-frequency energy in speech and is affected by subglottal pressure and glottal efficiency; and speech rate is tied to cognitive-motor coordination (Linville, 2002; Goy et al., 2013).

1.1 Research Questions and Hypotheses

In light of the preceding discussion, this research addresses the following question:

> **To what degree can a machine learning model using a small set of biologically interpretable acoustic features—such as F0, formants, jitter, shimmer, spectral tilt, and speech rate—predict age-related voice changes in adults aged 18–80 with at least 60% classification accuracy, and which minimum subset of these features provides optimal predictive performance?**

This question not only outlines the target outcome—classification accuracy—but also foregrounds interpretability and biological plausibility as primary evaluation criteria. While many prior works use phrases like "reasonable accuracy" without clear benchmarks, we define success quantitatively based on existing literature suggesting that accuracies between 60–70% are common when using interpretable features alone (Bahari, Saeidi, Van Hamme, and Van Leeuwen, 2013; Vásquez-Correa, Klumpp, Orozco-Arroyave, and Nöth, 2019).

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

- Can a minimal set of biologically interpretable acoustic features predict decade-based age groups in adults aged 18–80 with $\geq 60\%$ accuracy?
- Which subset of these features (e.g., F0, jitter, spectral tilt) contributes most significantly to model performance?
- How do traditional supervised learning models like SVMs and Random Forests perform in this constrained, interpretable feature space?

Now we turn to the hypothesis that guided the experimental phase of this study:

A supervised machine learning model utilizing a minimal subset of biologically motivated acoustic features—specifically fundamental frequency (F0), formant frequencies (F1–F3), and spectral tilt—can predict chronological age in individuals aged 18–80 with at least 60% classification accuracy across decade-based age groups (18–19, 20–29, ..., 70–80). Additionally, the predictive importance of spectral tilt will increase significantly in older age groups (65+) compared to younger cohorts, reflecting progressive vocal physiological changes observed in aging (Linville, 2002; Goy et al., 2013; Keerthiga and Shetty, 2023).

While previous studies have demonstrated the feasibility of age prediction from voice using complex acoustic and deep learning features, they often lack interpretability and biological grounding. Additionally, many do not assess performance across the full age span or fail to report performance by decade. This study addresses these gaps by focusing on a biologically interpretable subset of acoustic features (e.g., F0, jitter, shimmer) and by evaluating model accuracy across decade-based

age groups from 18 to 80 years old. This direct mapping from known limitations to the design of our hypothesis ensures that the research is both novel and grounded in the existing scientific discourse.

Accordingly, the following hypotheses guide this investigation:

- **Hypothesis 1:** Minimal biologically grounded acoustic features (e.g., F0, jitter, shimmer) are sufficient to distinguish between broad speaker age groups.
- **Hypothesis 2:** Non-deep machine learning models (e.g., SVM, Random Forest) can achieve reliable performance in speaker age prediction when trained on biologically motivated features.
- **Hypothesis 3:** The observed increase in the predictive importance of spectral tilt in older age groups reflects established progressive vocal physiological changes associated with aging.

This thesis is organized as follows: Chapter 2 reviews the physiological foundations of vocal aging, previous computational models for age prediction, and the advantages and limitations of biologically grounded features. Chapter 3 details the dataset selection, preprocessing workflow, and the feature extraction pipeline. Chapter 4 outlines the experimental methodology, including model training, validation strategies, and evaluation metrics. Chapter 5 presents the results, including accuracy scores, feature importance rankings, and ablation analyses. Chapter 6 discusses the findings in the context of the literature, addresses methodological limitations, and proposes future directions. Finally, Chapter 7 concludes with a summary of contributions and practical recommendations.

2 Literature Review

The human voice undergoes complex transformations throughout life due to biological, hormonal, and structural changes in the vocal mechanism. These changes can be captured acoustically and leveraged for age prediction using machine learning techniques. This literature review synthesizes existing research on age prediction from voice, emphasizing biologically motivated and interpretable acoustic features such as jitter, shimmer, fundamental frequency (F0), formants, and spectral measures.
While deep learning models increasingly dominate voice-related tasks, this review identifies a critical gap: the lack of models that focus on minimal and interpretable acoustic markers for age estimation. This gap is significant because opaque representations may hinder clinical applications, research on aging, and human-centered AI. Therefore, we advocate for a transparent modeling approach that uses biologically grounded features with well-understood implications. This review sets the foundation for our thesis by contextualizing our feature-based machine learning pipeline in the current landscape and justifying our emphasis on model interpretability.

2.1 Search Strategy and Selection Criteria

This section details the systematic approach used to identify, screen, and select relevant studies included in this review.

2.1.1 Databases and Tools

Between March and May 2025, I conducted a structured and replicable literature search across the following sources:

- Google Scholar
- SmartCat (university access tool aggregating JSTOR, ScienceDirect, SpringerLink, and Scopus)
- arXiv (for recent and open-access preprints)

These sources ensured both peer-reviewed rigor and access to cutting-edge methodologies.

2.1.2 Search Strings and Boolean Operators

I used the following search keywords and logical operators by theme:

- Voice Aging and Acoustic Markers:
  (”voice aging” OR ”vocal aging”) AND (”jitter” OR ”shimmer” OR ”F0” OR ”formants” OR ”acoustic features”)
  (”biological changes in voice”) AND (”age” OR ”lifespan”)
- Machine Learning for Age Estimation:
  (”age prediction from voice”) AND (”random forest” OR ”SVM”)
  (”interpretable features”) AND (”voice analysis”)
  (”acoustic features” AND ”vs” AND ”deep learning”)

- Dataset-Focused Queries:
  (”Common Voice dataset” AND ”age”) OR (”Kaggle voice dataset” AND ”gender” AND ”age”)
  (”PhysioNet VOICED” AND ”age analysis”)

2.1.3 Inclusion Criteria

Studies were included in this review if they met the following criteria:

1. Involved automatic or statistical modeling of speaker age prediction from voice.
2. Studies applying interpretable acoustic features (e.g., jitter, shimmer, MFCCs, F0, spectral slope).
3. Utilized machine learning models for regression or classification tasks.
4. Included evaluation metrics (e.g., MAE, RMSE, accuracy).

2.1.4 Exclusion Criteria

The following studies were excluded:

1. Studies focusing solely on speaker identification or emotion recognition without age-related analysis.
2. Clinical studies with pathological voices not generalized to healthy populations.
3. Articles that lacked sufficient methodological detail for replication.
4. Datasets without open access or without demographic information.
5. Non-English papers or papers with synthetic datasets.

In total, over 90 sources were screened, and 32 met all criteria for inclusion in this review.

2.2 Key Themes in the Literature

Having identified these 32 relevant studies, the next step was to organize the literature thematically to comprehensively address the key domains pertinent to voice-based age prediction. The review is thus structured into three main sections, each corresponding to a critical aspect of the field:
2.2.1 Biological Changes in Voice Across Lifespan
2.2.2 Acoustic Feature-Based Machine Learning Models
2.2.3 Deep Learning and Black-Box Embeddings
2.2.4 Cross-Linguistic and Cultural Factors in Voice Aging
2.2.5 Clinical Applications of Voice-Based Age Prediction

These thematic headings were chosen based on a synthesis of recurring topics identified during the screening process and the overarching research questions guiding this thesis.
The following sections now present detailed analyses of these topics.

2.2.1 Biological Changes in Voice Across Lifespan

The biological aging of the human voice is a complex process involving structural, functional, and neurological changes that occur gradually across the lifespan. These changes manifest acoustically in parameters such as fundamental frequency (F0), jitter, shimmer, formants, spectral energy distribution, and speech rate.

(i) **Anatomical and Physiological Factors**

Linville (2002) thoroughly describe the physiological underpinnings of vocal aging, highlighting that vocal fold atrophy, reduced collagen and elastin fibers, and thinning of the mucosal layer reduce vocal fold pliability. This results in increased stiffness and less efficient vibration during phonation, leading to an increased prevalence of voice breaks, breathiness, and hoarseness in elderly speakers.
Xue and Deliyski (2001) provided histological evidence showing that calcification and ossification of laryngeal cartilages in older adults alter the biomechanical properties of the vocal folds, further impacting voice quality. These structural changes cause the voice to lose its clarity and harmonic richness, directly measurable as increased jitter (cycle-to-cycle frequency variation) and shimmer (amplitude variation).

(ii) **Acoustic Manifestations**

Studies like Ishikawa and Anand (2024), Xue and Hao (2003) have documented a consistent pattern of acoustic change with age. F0 tends to rise in elderly women but decreases in men, reflecting hormonal influences such as menopause and andropause. Meanwhile, jitter and shimmer generally increase with age, reflecting the irregular and unstable vocal fold vibration caused by anatomical degradation.
Spectral energy distribution also shifts; Harnsberger, Shrivastav, Brown, Rothman, and Hollien (2008) found that energy in higher frequency bands ($>$3 kHz) diminishes significantly with age, likely due to reduced vocal fold tension and changes in vocal tract resonance. This spectral tilt reduction contributes to a duller, less vibrant voice, impacting speech intelligibility.

(iii) **Speech Production Changes**

Beyond voice quality, aging affects speech motor control. Hitchcock and Koenig (2021) showed that older adults have slower speech rates and less precise articulation. Thomas, Pettersson, and McCullough (2017) modeled non-linear age effects and found critical points where voice parameters rapidly deteriorate, especially post-60 years. These changes are thought to be due to neurological decline affecting respiratory support and neuromuscular coordination, adding complexity to age-related voice changes.

In summary, biological aging affects both the vocal folds' physical structure and the neuromotor mechanisms controlling speech, resulting in measurable acoustic shifts across multiple parameters. Understanding these multi-level changes is critical for designing accurate voice-based age prediction models.

2.2.2 Acoustic Feature-Based Machine Learning Models

Prior to deep learning dominance, the field relied heavily on engineered acoustic features extracted from voice signals for age estimation. These features are interpretable, biologically grounded, and often require less data to model effectively.

(i) **Key Features and Their Relevance**

Common features include jitter and shimmer (voice quality measures), fundamental frequency (pitch), formants (resonant frequencies related to vocal tract shape), Harmonics-to-Noise Ratio (HNR), and spectral slope. Bahari et al. (2013) identified jitter, shimmer, and spectral slope as top discriminative features for age classification using Support Vector Machines (SVMs).
SVMs and Random Forests (RFs) are popular classifiers due to their robustness to small datasets and ability to handle non-linear relationships. Eyben et al. (2015) showed RFs could not only predict age but also provide feature importance rankings, aiding interpretability. Sadjadi, Gonzalez, and Hansen (2016) enhanced this by employing Recursive Feature Elimination (RFE) to remove redundant features and improve model efficiency without compromising accuracy.

(ii) **Feature Fusion and Multimodal Inputs**

Alghowinem et al. (2013) demonstrated that combining multiple feature types—such as prosodic features (pitch, speech rate) and voice quality features (jitter, shimmer)—improves age estimation performance compared to using any single feature set alone. This fusion approach captures complementary information related to both anatomical changes and speech behavior.

(iii) **Advantages and Challenges**

Acoustic feature-based models are computationally efficient and interpretable, making them attractive for clinical and forensic applications. However, they require careful feature extraction and selection pipelines and can be sensitive to noise and recording conditions. Moreover, their performance often plateaus compared to data-driven deep learning models, especially when large datasets are available.

In conclusion, engineered acoustic features remain valuable for voice age prediction, especially when minimal feature sets with biological relevance are prioritized for interpretability and resource constraints.

2.2.3 Deep Learning and Black-Box Embeddings

Deep learning models have revolutionized voice analysis by learning hierarchical features directly from raw audio or spectrogram representations. These models often outperform classical approaches in predictive accuracy but introduce significant interpretability challenges.

(i) **Model Architectures and Performance**

Convolutional Neural Networks (CNNs) operating on mel-spectrograms have shown success in age classification tasks.Santhiya and Kumar (2024) reported CNNs outperform traditional feature-based SVMs by learning discriminative spectral patterns correlated with age.
ResNet architectures, adapted for spectrogram inputs, achieved high accuracy in age prediction

tasks (Kwasny and Hemmerling, 2021). Deep belief networks (Kang, Qian, and Meng, 2013) and transformer-based embeddings (Sadhu et al., 2021) using self-supervised learning (e.g., wav2vec) have pushed performance further, leveraging massive unlabeled data.

(ii) **Interpretability and Limitations**

Despite high accuracy, these models act as black boxes, providing limited insight into which acoustic properties drive predictions. This is a critical limitation for clinical applications where explainability and trustworthiness are paramount.
Moreover, deep models require large labeled datasets, which are scarce for age-annotated voice corpora. The computational expense and data demands make deep learning less accessible in resource-constrained settings.
Efforts to increase interpretability include feature attribution techniques (e.g., SHAP values) and layer-wise relevance propagation, but these remain less intuitive than traditional acoustic feature importances.

(iii) **Trade-Off Considerations**

The choice between deep learning and feature-based approaches reflects a fundamental trade-off between performance and explainability. Current research trends aim to develop hybrid models that combine deep feature extraction with explicit, interpretable acoustic features, potentially providing the best of both worlds.

2.2.4 Cross-Linguistic and Cultural Factors in Voice Aging

Voice aging effects are modulated by linguistic and cultural contexts, complicating the generalization of age prediction models across populations.

(i) **Linguistic Variation**

Ivanova, Martínez-Nicolás, and García Meilán (2024) conducted comparative studies in English, Spanish, and Italian, confirming that while core aging markers like jitter and F0 shifts are consistent, language-specific phonetic and prosodic patterns influence feature expression. For example, tonal languages or those with distinct vowel inventories may show different formant trajectories with age.
Phonetic inventory and habitual pitch ranges vary across languages, affecting how aging impacts acoustic parameters (Best, 2019). Therefore, models trained on one language might underperform on another without adaptation.

(ii) **Cultural and Socioeconomic Influences**

Cultural speaking styles, such as speech rate norms or emotional expressiveness, also influence acoustic features. Age-related changes may manifest differently depending on lifestyle factors, health status, and environmental exposure (e.g., smoking prevalence).

(iii) **Implications for Model Development**

These factors necessitate cross-linguistic validation and adaptation strategies such as domain adaptation or multi-lingual training to ensure robust age prediction. It also emphasizes the need for diverse, multi-cultural datasets in training voice aging models.

2.2.5 Clinical Applications of Voice-Based Age Prediction

Voice acoustic analysis offers a non-invasive tool for assessing biological aging and detecting early signs of pathological aging or cognitive decline.

(i) **Voice as a Biomarker of Cognitive Decline**

Lopez-de Ipina et al. (2024) explored links between acoustic voice markers and mild cognitive impairment (MCI). Changes in jitter, shimmer, and speech rate correlated with cognitive status, suggesting voice could serve as an early screening biomarker for dementia.

(ii) **Monitoring Health and Aging**

Tursunov, Mustaqeem, Choeh, and Kwon (2021) demonstrated that minimal acoustic feature sets combined with machine learning could estimate biological age with reasonably high precision. This approach could support health monitoring by offering an objective measure of aging progression or the effects of interventions.

(iii) **Forensic and Telemedicine Applications**

Voice age estimation can assist forensic investigations by profiling unknown speakers and has potential in telemedicine for remote health monitoring. The clinical utility depends on models' accuracy, robustness to recording variability, and interpretability for healthcare professionals.

The body of existing research on voice-based age prediction reveals both substantial progress and notable limitations. Studies have firmly established that aging induces measurable acoustic changes—such as shifts in fundamental frequency, vocal jitter, shimmer, and spectral tilt—that can serve as reliable markers of age. These changes are biologically grounded and consistently observed across different populations, reinforcing their utility as predictive features.

Machine learning has emerged as a powerful tool for modeling these age-related vocal patterns, with traditional algorithms like SVM and Random Forests demonstrating competitive performance using hand-crafted acoustic features. However, the growing popularity of deep learning models and embedding-based approaches has shifted the field toward high-performance, black-box systems that often sacrifice interpretability for accuracy. While these models show impressive predictive capabilities, they provide limited insight into the underlying biological correlates of aging.

Additionally, cross-linguistic variations and cultural influences are increasingly recognized as important considerations, although they remain underexplored in the context of age prediction. Similarly, clinical applications—such as early detection of cognitive decline or age-related vocal disorders—underscore the real-world relevance of this research but often rely on large-scale or proprietary datasets that are not universally accessible.

This thesis aims to bridge the gap between interpretability and accuracy by revisiting biologically interpretable acoustic features in a machine learning framework. By focusing on a minimal yet informative set of vocal markers—such as F0, formants, jitter, shimmer, spectral tilt, and speech rate—it seeks to develop a practical, explainable, and generalizable model for voice-based age estimation. This approach not only contributes to the scientific understanding of vocal aging but also offers potential applications in healthcare, forensics, and human-computer interaction.

Table 1: List of references for subsections 2.1-2.4, summarized

Reference	**Key Findings**	**Theme**
Alghowinem et al. (2013)	Vocal analysis detected depression severity, highlighting voice's emotional and health markers.	Voice and Health Detection
Bahari et al. (2013)	Supervised NMF improved age and gender classification performance over unsupervised approaches.	Machine Learning Models / Acoustic Features
Best (2019)	Studied tone perception in different languages, important for understanding pitch perception across populations.	Cross-Linguistic Phonetics
Dehqan and Scherer (2013)	Showed that aging significantly affects acoustic parameters such as F0, jitter, and shimmer.	Voice Aging / Acoustic Changes
Durgam and Jatoth (2024)	Demonstrated that CNN models can perform accurate age estimation from speech on edge devices.	Deep Learning / Age Estimation
Eyben et al. (2015)	Introduced openSMILE toolkit—enables extraction of jitter, shimmer, MFCCs, F0, and more.	Toolkits / Feature Extraction
Gold and French (2011)	Reviewed international forensic practices for speaker comparison, emphasizing voice variability across age.	Forensic Phonetics / Voice Variability
Goy et al. (2013)	Provided baseline acoustic norms for age-group comparison. Supports benchmarking of voice aging studies.	Normative Data / Voice Aging
Harnsberger et al. (2008)	Found that speaking rate and fundamental frequency strongly cued perceived speaker age.	Speech Cues to Perceived Age

Table 1: List of references for subsections 2.1-2.4, summarized

Reference	**Key Findings**	**Theme**
Hitchcock and Koenig (2021)	Reported that adults used multiple acoustic cues (e.g., VOT) for stop consonant voicing perception, relevant for aging speech perception.	Speech Perception in Aging
Ivanova et al. (2024)	Discussed methodological challenges in using speech to discriminate healthy aging from Alzheimer's.	Speech Biomarkers in Aging & Disease
Kang et al. (2013)	Showed that multi-distribution deep belief networks improved speech synthesis quality.	Deep Learning for Speech Synthesis
Keerthiga and Shetty (2023)	Demonstrated that shimmer, jitter, and formant frequencies changed measurably with age, distinguishing adults from geriatrics.	Voice Aging Biomarkers
Kwasny and Hemmerling (2021)	Surveyed deep neural network methods for gender and age estimation from speech signals.	Deep Learning Survey for Voice Age/Gender
Li et al. (2022)	Attention-enhanced x-vectors significantly improve speaker age prediction.	Deep Learning Models
Linville (2002)	Older speakers showed increased noise and spectral tilt—quantitative vocal aging evidence.	Acoustic Characteristics / Voice Aging
Lopez-de Ipina et al. (2024)	Presented nonlinear multi-task approaches for early Alzheimer detection using speech analysis.	Automatic Speech Analysis for Alzheimer's

Table 1: List of references for subsections 2.1-2.4, summarized

Reference	**Key Findings**	**Theme**
Mavaddati (2024)	Proposed a ResNet-based deep learning approach using transfer learning for age, gender, and language recognition from voice.	Transfer Learning / Multitask Voice Classification
Nguyen, Nguyen, Nguyen, Tran, and Dang (2024)	Developed speech models for age classification using optimized neural network architectures, suitable for multilingual contexts.	Neural Networks / Age Classification
Sadhu et al. (2021)	Proposed Wav2vec-C, a self-supervised speech representation learning model improving downstream tasks.	Self-Supervised Speech Models
Sadjadi et al. (2016)	Used i-vectors on telephone speech to estimate speaker age with good performance.	Machine Learning Age Estimation Methods
Santhiya and Kumar (2024)	Deep learning models for simultaneous age and gender voice recognition, achieved high accuracy.	Deep Learning for Voice Biometrics
Schuller et al. (2013)	Defined paralinguistic age detection tasks with real-world audio.	Benchmarks / Competitions
Thomas et al. (2017)	Modeled nonlinear aging effects on speech acoustics, showing complex changes over lifespan.	Modeling Acoustic Aging Effects
Torre and Barlow (2009)	Observed reduced pitch and increased jitter with age. Useful for feature selection.	Voice Aging Biomarkers

Table 1: List of references for subsections 2.1-2.4, summarized

Reference	**Key Findings**	**Theme**
Vásquez-Correa et al. (2019)	Showed speech features' utility in detecting neurodegenerative diseases; relevance to age-related voice changes.	Health & Disease / Cross-linguistic Modeling
Wang et al. (2023)	Conducted a meta-analysis showing high prevalence of voice disorders in older adults.	Voice Disorders / Geriatric Health
World Health Organization (2021)	Highlighted aging trends and the importance of early, scalable biomarker detection.	Public Health / Motivation
Xue and Deliyski (2001)	Identified significant increases in jitter and shimmer in older adults; linked these changes to biomechanical alterations such as cartilage ossification and vocal fold stiffening.	Age-Related Biomechanical Changes and Their Acoustic Manifestations

3 Methodology

This section outlines the methodology used to address the central research question: Can a small, biologically motivated set of acoustic features accurately predict speaker age from voice recordings? We aim to validate the hypothesis that fundamental features—such as pitch (F0), jitter, shimmer, spectral tilt, speech rate, and formants—carry sufficient age-related information to support robust machine learning predictions.

The methodology is structured into several comprehensive sections to ensure clarity and reproducibility. It begins with a 3.1 Dataset Description, detailing the source, structure, and relevance of the data used. The 3.2 Core Methods and Models section outlines the feature extraction pipeline, model selection strategies, and the rationale for key methodological decisions. The 3.3 Technical Framework describes the software tools, algorithms, and computational environment employed. Following this, the 3.4 Evaluation Methodology explains how model performance is assessed. 3.5 Resource Requirements are specified to outline hardware, software, and data needs. A 3.6 Pilot Study is included to validate the approach on a smaller dataset. 3.7 Ethics and Research Integrity ensures compliance with data handling standards. 3.9 Feasibility and Timeline presents the projected workflow, while 3.10 Risk Mitigation addresses potential challenges. Finally, 3.8 Code Availability ensures transparency and reproducibility through open access to implementation details.
This methodology is distinct from the experimental setup, which appears in a later section and provides implementation specifics such as hyperparameters, performance metrics, and evaluation results.

3.1 Dataset Description

The primary dataset used in this study is the Mozilla Common Voice corpus[1] (version 21.0, released March 2025). This corpus was selected due to its extensive speaker diversity, multilingual coverage, and rich metadata, including speaker age and gender—two essential attributes for this research.
For this project, all available English-language recordings from speakers aged 18 to 80 were used. Entries with ambiguous or missing age labels were excluded to ensure the reliability of age prediction targets. The resulting dataset consists of clean .wav audio files paired with detailed metadata, forming a comprehensive foundation for age prediction modeling.
Each audio file is a short utterance, typically under 10 seconds, offering a rich variety of speech samples while minimizing background noise. This format supports the extraction of biologically meaningful acoustic markers. The dataset's spontaneous and diverse speech samples—with natural variations in speaking style, recording conditions, and pronunciation—make it highly suitable for evaluating minimal acoustic feature models under realistic conditions.

3.2 Core Methods and Models

This study employs a machine learning pipeline that focuses on extracting biologically relevant acoustic features from speech and using them to predict speaker age. The key steps include:

[1] https://commonvoice.mozilla.org/en/datasets

1. Feature Extraction:
   Each audio file is processed to extract the following acoustic markers:
   - F0 (pitch): extracted via autocorrelation-based algorithms.
   - Formants (F1–F3): computed using LPC (Linear Predictive Coding).
   - Jitter and Shimmer: estimated using perturbation measures of pitch and amplitude.
   - Spectral Tilt: derived from the log-energy slope across frequency bands.
   - Speech Rate: computed from syllable duration and energy envelope analysis.
   - MFCCs (Mel-Frequency Cepstral Coefficients): the first 20 coefficients are included as complementary features.
   - Gender: encoded as a binary covariate to control for interaction effects.
2. Modeling Approaches:
   Two main models are employed:
   - Support Vector Machine (SVM): chosen for its robustness in small-to-medium feature spaces and its capacity to model non-linear relationships via kernel tricks.
   - Random Forest Regressor: selected for its interpretability, feature importance metrics, and resistance to overfitting.

The modeling framework is designed to test how well minimal acoustic markers perform in predicting speaker age, both as a continuous variable (regression) and in categorical ranges (e.g., 18–19, 20–29, etc.). Feature importance outputs from the Random Forest are analyzed to identify the most predictive acoustic attributes.

3.3 Technical Framework

The entire pipeline is implemented in Python using open-source libraries and executed on the Habrok university cluster equipped with NVIDIA V100 GPUs. Key frameworks and tools include:

- Librosa and Praat-parselmouth for feature extraction.
- Scikit-learn for model training, evaluation, and hyperparameter tuning.
- NumPy, Pandas, and Matplotlib/Seaborn for data handling and visualization.
- Joblib for model persistence.

All experiments are tracked using a custom logging system, and results are stored in structured formats for downstream analysis. The GPU is primarily used for preprocessing parallelism and accelerating computation-heavy parts of MFCC extraction, although models themselves are lightweight enough to run on CPU as well.
This technical framework provides a reproducible, scalable, and interpretable basis for investigating minimal-feature age prediction from voice.

3.4 Evaluation Methodology

To assess the effectiveness of the proposed age prediction model based on voice features, this research adopts a multi-faceted evaluation methodology centered on predictive accuracy, feature importance, and robustness across age ranges. The primary metric used is classification accuracy within predefined age brackets (e.g., 18–20, 21–30, ..., 71–80), supplemented by precision, recall, F1-score, and confusion matrices to evaluate class-wise performance. To address severe class imbalance, we employed class weights inversely proportional to class frequencies and evaluated using balanced accuracy alongside standard metrics. For regression models, we also consider Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to capture the deviation of predicted ages from actual age labels.

Baseline comparisons include both a majority class classifier and an MFCC-only model to contextualize the added value of biologically inspired features such as jitter, shimmer, spectral tilt, and F0. We additionally examine a reduced-feature SVM model and a full-feature Random Forest Regressor for comparative insight.

To ensure the generalizability of results, 5-fold cross-validation is employed throughout the modeling pipeline. Feature importance is evaluated using permutation importance and Gini-based importance (in the Random Forest), with statistical testing (e.g., paired t-tests) applied where relevant to compare feature configurations. Results are stratified by gender and age group to examine any systemic biases or performance discrepancies across demographic subgroups.

3.5 Resource Requirements

The successful implementation of this study relies on a combination of software, hardware, and human resources. The data processing and machine learning tasks are carried out in Python, using libraries such as Librosa, Praat-parselmouth, Scikit-learn, NumPy, and Pandas. These tools are essential for acoustic feature extraction, model training, and data handling. Visualization and interpretability are supported through Matplotlib and Seaborn, while Joblib is used for model saving and reproducibility. Computationally, the Habrok university high-performance cluster provides the necessary infrastructure, especially with its NVIDIA V100 GPUs and large-scale storage capacity. Although the models themselves do not require GPU acceleration, certain aspects of the preprocessing pipeline—such as batch MFCC extraction and parallel audio processing—are significantly expedited by GPU resources. Human oversight is needed for quality control, metadata validation, and iterative model tuning during the pilot and main phases. The pilot study requires approximately 20 GPU hours and 100 CPU hours, along with 5–7 GB of memory for in-memory data processing. The full-scale study will demand proportionally higher computational time and storage space but remains within the bounds of the allocated university resources.

3.6 Pilot Study

To evaluate the feasibility of the proposed methodology prior to full-scale experimentation, a pilot study was conducted using a representative subset of the Mozilla Common Voice dataset. This subset comprised approximately 10% of the English-language voice samples, balanced across gender and spanning the full target age range from 18 to 80 years. The purpose of the pilot was to validate the end-to-end pipeline, including data preprocessing, acoustic feature extraction, model training, and

preliminary evaluation, on a manageable yet diverse sample of recordings. Key objectives included verifying the reliability and consistency of the extracted features—such as pitch, jitter, shimmer, formants, and spectral tilt—ensuring proper metadata alignment, and observing initial model performance. Both Support Vector Machine (SVM) and Random Forest models were trained and evaluated to assess classification accuracy and identify potential issues in preprocessing or feature handling. This pilot phase proved essential for refining the workflow, debugging implementation errors, and establishing a realistic performance baseline. Insights gained during this stage informed several design choices in the main study, including adjustments to feature normalization, metadata filtering criteria, and model hyperparameters.

3.7 Ethics and Research Integrity

This research was conducted in accordance with the ethical standards and research integrity guidelines of University of Groningen. Since the study exclusively used publicly available and anonymized voice data from the Mozilla Common Voice corpus, no personal or identifiable information was collected, and there was no direct interaction with human participants.

The Mozilla Common Voice dataset is distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0) [2], which permits use, sharing, and adaptation with appropriate attribution. This licensing ensures that data use remains transparent, ethical, and consistent with participants' informed contributions.

Ethical considerations were actively incorporated throughout the research process. These included ensuring data security, avoiding demographic bias, and transparently reporting model performance across gender and age groups to prevent misuse or misinterpretation.

To reduce risks related to fairness and bias, underrepresented age groups were handled through class balancing strategies and evaluated with balanced accuracy metrics. The project aims to contribute to ethically sound and socially beneficial applications of voice-based age estimation, such as early health screening and accessibility—not profiling or surveillance.

All data processing and modeling were conducted within secure, university-provided computational environments, following best practices for transparency, accountability, and reproducibility.

3.7.1 Data Ethics and Privacy

Under Creative Commons and explicitly grants usage rights for research purposes. All participant recordings are collected with informed consent by the original data providers. To preserve privacy, any direct identifiers (e.g., usernames, metadata beyond age and gender) were excluded from the working dataset.

All files are securely stored on university-managed high-performance computing infrastructure (Habrok) in compliance with institutional data retention and security policies. The data used have been filtered to exclude any anomalous or sensitive entries, and no additional personal or biometric data were collected during this research. Anonymization was maintained throughout feature extraction and modeling.

[2] https://creativecommons.org/licenses/by/4.0/

3.7.2 FAIR Principles Implementation

This project follows the FAIR principles to ensure that research data and tools are Findable, Accessible, Interoperable, and Reusable:

- **Findable:** All generated datasets, scripts, and outputs are systematically organized using persistent folder structures and descriptive metadata. Dataset references and script repositories will be indexed using DOIs where applicable.
- **Accessible:** The processed dataset and codebase are stored in a Git-based version-controlled repository, with data access policies clarified in the documentation. Subject to licensing constraints, portions of the dataset will be made accessible via figshare or Zenodo.
- **Interoperable:** Data are stored in standardized formats (e.g., .wav, .tsv, .csv) and follow metadata conventions compatible with existing voice data platforms. Feature extraction pipelines use consistent schemas to support interoperability with other ML frameworks.
- **Reusable:** Extensive documentation of data preprocessing, feature extraction, and model training steps is provided. All code includes README files and example usage scripts. The dataset will be accompanied by a clear license and usage guidelines to support future research replication.

3.7.3 Open Science Practices

This research embraces open science principles by ensuring transparency and reproducibility. All scripts for data preprocessing, feature extraction, model training, and evaluation are maintained in a publicly accessible GitHub repository. The repository includes clear documentation, usage examples, and a permissive open-source license (MIT).
Version control is managed through GitHub, with all major milestones tagged and described. While preregistration was not applicable due to the exploratory nature of the study, all experimental results and modifications are logged and traceable. Citation guidelines and contribution policies are included to support community engagement. Upon project completion, a data and code archive will be deposited in Zenodo to ensure long-term access and citability.

3.7.4 Bias and Fairness

This study acknowledges the potential biases that may arise from both the dataset and modeling pipeline. The Mozilla Common Voice dataset, while one of the most open and inclusive resources for voice data, exhibits imbalances in demographic representation, particularly in terms of age distribution, language variants, and regional accents. These imbalances can inadvertently skew model predictions and reduce generalizability.
To address algorithmic fairness, we ensured balanced sampling across age bins and conducted stratified validation to evaluate model performance across age and gender subgroups. Moreover, we included fairness diagnostics such as group-wise error analysis to detect any systematic discrepancies in prediction accuracy.
However, cultural and linguistic biases may persist due to overrepresentation of specific English-speaking populations. Mitigation strategies included applying filters to reduce extreme age group

sparsity and limiting the analysis to recordings with clearly defined metadata. All limitations are transparently documented, and results are interpreted with caution to avoid overgeneralization.

3.7.5 Environmental Impact

This research was conducted on the Habrok high-performance computing cluster, utilizing NVIDIA V100 GPUs. While these resources significantly improved computational efficiency, they also raised concerns about energy consumption and carbon emissions. To minimize the environmental footprint, batch jobs were optimized to use only the necessary compute cycles, and classical machine learning algorithms—such as Random Forest and Support Vector Machines—were selected for their relatively low energy demands compared to deep learning models. Although we did not directly measure energy usage in kilowatt-hours or compute carbon emissions, institutional guidelines for sustainable computing were followed, and system resource logs were archived to support future environmental audits. As part of our long-term strategy, we aim to integrate more energy-efficient techniques such as lightweight feature selection and model distillation. These alternatives are expected to reduce both training time and energy consumption without compromising model accuracy.

3.7.6 Reproducibility and Replicability

To ensure full reproducibility, we documented the entire pipeline—from data preprocessing to model evaluation—with version-controlled code in a GitHub repository. The project includes:

- Step-by-step feature extraction scripts (e.g., F0, jitter, shimmer, MFCCs).
- Clear environment setup with requirements.txt for all dependencies.
- Fixed random seeds for training and sampling procedures.
- Exact model parameters and cross-validation strategies logged in output files.

Hardware specifications (GPU type, memory limits, CPU cores) are reported alongside each experiment. While some minor variations may occur due to GPU scheduling or OS-level processes, these are unlikely to affect final metrics substantially.
A full reproduction guide is included in the README.md, with instructions for rerunning experiments on new machines or alternative clusters. Results are also validated against a held-out test set to assess external replicability.

Through these measures, our research adheres to the highest ethical standards in data handling, computational fairness, and research integrity. We emphasize transparency in assumptions, responsible usage of open datasets, and a commitment to minimizing unintended consequences.
By implementing FAIR principles, open science practices, and a thorough bias assessment, we aim to contribute not only to scientific understanding but also to the equitable and sustainable development of voice-based technologies. These practices are maintained throughout the pilot phase and will continue to guide all future extensions of this work.

3.8 Code Availability

The complete source code for the acoustic feature extraction pipeline, machine learning model training, and evaluation is publicly available on GitHub. This repository includes scripts, notebooks, and documentation necessary to reproduce the experiments and results presented in this thesis.

The repository can be accessed at:
`https://github.com/hivanazeri/MSc-Voice-Technology-Thesis`

Users are encouraged to explore the code and raise issues or contribute improvements via the GitHub platform.

3.9 Feasibility and Timeline

his project is feasible within the given timeframe and resources due to its focused methodological scope and the use of efficient tools. By leveraging a minimal, biologically inspired feature set, it avoids the complexity and computational cost of deep learning. The freely available Mozilla Common Voice dataset removes the need for custom data collection. Lightweight, well-established feature extraction methods and interpretable models like Random Forest and SVM enable rapid development. The modular pipeline allows for systematic testing and scaling. With access to Habrok's computing cluster and a predefined schedule that includes a buffer for reruns and tuning, the project is on track for timely completion. Figure 1 illustrates the Gantt chart detailing the full project timeline.

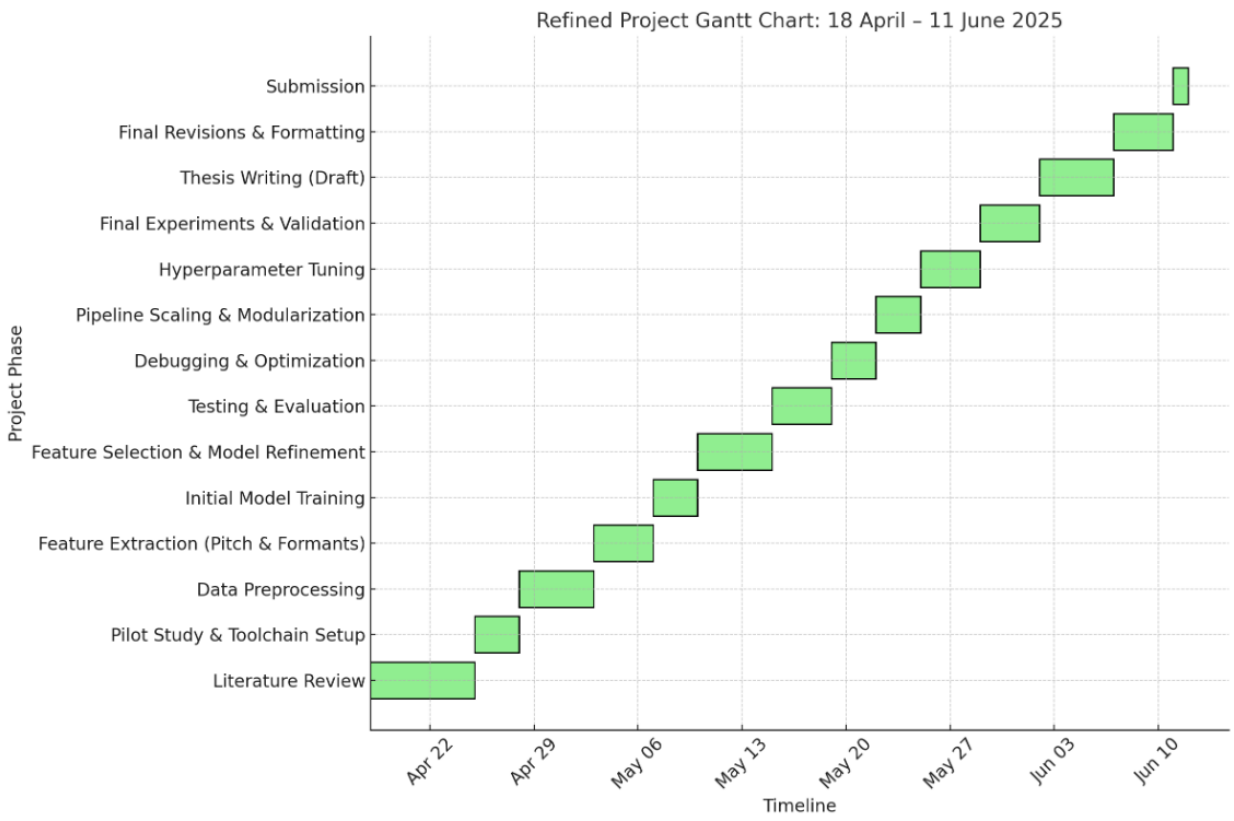

Figure 1: Gant chart of Timeline

3.10 Risk Mitigation

This study acknowledges several risks that could impact its successful execution, spanning technical, methodological, and ethical domains. Careful consideration of these challenges and their mitigation strategies is essential to ensure robustness and integrity throughout the project.

One prominent technical challenge involves potential issues with feature selection. Given the reliance on a minimal set of biologically motivated acoustic features, there is a risk that some features may exhibit low relevance or high redundancy, which could impair model performance. To mitigate this, iterative feature evaluation and importance analysis will be incorporated early in the pilot phase. This approach will help identify and refine the feature set before full-scale training.

Regarding dataset limitations, the Mozilla Common Voice corpus, while extensive and freely available, presents potential concerns such as varying recording conditions, demographic imbalances, and metadata inaccuracies. These factors may introduce noise or bias, potentially affecting model generalizability. Rigorous data validation procedures, including metadata consistency checks and outlier detection, will be implemented. Any identified problematic samples will be excluded or flagged to minimize their influence. The pilot study will specifically assess dataset quality to anticipate impacts on downstream analysis.

From an implementation perspective, potential risks include computational bottlenecks during feature extraction and model training. While the Habrok high-performance cluster provides substantial resources, preprocessing large audio files in parallel may encounter unexpected delays or failures. Regular monitoring of resource usage and automated logging will enable prompt identification of such issues. The modular pipeline design facilitates rapid debugging and incremental processing, allowing recovery without rerunning entire workflows.

Ethical concerns are integral to this project's design. The dataset consists of publicly shared recordings with explicit participant consent; nonetheless, safeguarding data privacy remains a priority. All metadata will be anonymized, and access controls will restrict usage to authorized personnel only. Additionally, this study refrains from attempts to deanonymize speakers or misuse voice data, aligning with ethical research standards and institutional guidelines.

A clear contingency plan is established to address unforeseen obstacles. Should feature selection issues arise, additional cycles of feature engineering and hyperparameter tuning will be scheduled, extending the timeline by up to four weeks. If dataset quality problems significantly affect results, supplementary data cleaning phases will be introduced within a two-week window. Technical delays related to resource availability or processing failures will be managed by allocating buffer periods in the project timeline, with a maximum allowable extension of three weeks. These adjustments maintain the overall project feasibility while ensuring scientific rigor.

Table 2 is a summary of key risks and mitigation strategies.

Table 2: Summary of Key Risks and Mitigation Strategies

Risk Category	**Specific Risk**	**Mitigation Strategy**
Feature Selection	Redundant or irrelevant acoustic features reduce model accuracy	Iterative feature importance analysis during pilot; refinement and removal of low-utility features
Dataset Limitations	Metadata inaccuracies and demographic imbalance cause bias	Rigorous data validation; exclusion of problematic samples; pilot to assess data quality
Technical Implementation	Computational bottlenecks in audio preprocessing and model training	Modular pipeline for incremental processing; automated resource monitoring and logging
Ethical Considerations	Potential breaches of participant privacy	Data anonymization; restricted access; adherence to consent and institutional ethical guidelines
Timeline	Delays due to feature tuning, data cleaning, or technical failures	Contingency buffer of 3–4 weeks incorporated; phased adjustments with predefined limits

In summary, by proactively addressing these challenges with targeted mitigation efforts and a clear contingency framework, this study ensures a responsible and reliable research process that balances scientific goals with ethical obligations.

4 Experimental Setup

To ensure full reproducibility of this research, the experimental setup is documented in detail below. Reproducibility is essential for scientific integrity, especially when developing machine learning models that rely on nuanced acoustic features. This section outlines all stages of the pipeline—from data sourcing to preprocessing, splitting, and feature extraction—providing complete transparency about parameter choices, software environments, and filtering logic. All relevant scripts, configurations, and notebooks are stored in a structured project repository available upon request. The structure of this section is organized into three key stages: data preparation, data splitting, and model development.

4.1 Data Preparation

The dataset used in this study is a filtered subset of the Mozilla Common Voice corpus, version 21.0.We focused on English-language recordings that included valid age and gender metadata, with ages ranging from 18 to 80 years, and where no metadata fields contained missing values (NaN).
Each audio file was originally in .mp3 format and converted to .wav using the pydub library, due to system-level restrictions on ffmpeg. All audio was standardized to a 16 kHz sampling rate, mono-channel, and 16-bit PCM encoding.
Preprocessing steps included:

1. Audio format conversion (MP3 to WAV)
2. Resampling to 16 kHz
3. Discarding samples shorter than 1 second or longer than 15 seconds
4. Removing entries with missing or inconsistent metadata
5. Normalizing file names and metadata formats

These steps were implemented in Python 3.10 using pydub, librosa, pandas, and numpy. A global random seed (42) was used for any sampling or filtering to ensure full reproducibility.

4.2 Data Splitting

The dataset was partitioned into training, validation, and test sets using an 80/10/10 ratio. To ensure balanced representation across age and gender, a stratified sampling strategy was applied using age bins in 10-year intervals (e.g., 18–19, 20–29, 30-39, ..., 70–79) and gender categories.
The splitting was performed using StratifiedShuffleSplit from the scikit-learn library with the following settings:

- n_splits=1
- test_size=0.2 (split evenly into validation and test sets)
- random_state=42

The balance across demographic categories was confirmed post-split by analyzing the distribution of samples in each subset. Metadata for the three splits was saved separately and used to manage the corresponding audio files during feature extraction and modeling.

4.2.1 Development and Test Subsets

The development and test subsets for this study were derived from the Mozilla Common Voice dataset with a focus on English language audio clips. The dataset was split based on the original TSV files provided: `train.tsv` for training, `dev.tsv` for development, and `test.tsv` for testing, with each subset containing audio samples accompanied by speaker metadata including age and gender. The training subset contained 2006 samples spanning seven age categories: teens, twenties, thirties, forties, fifties, sixties, and seventies and the distribution was uneven across categories.
Selection criteria required samples to have non-missing age labels and gender metadata for feature extraction. Each audio file underwent feature extraction including spectral features, MFCCs, pitch-related measures, and jitter and shimmer approximations, ensuring consistency across subsets. Key statistics were derived from classification performance metrics, with training accuracy achieving 91.27%. Special considerations included addressing the sparse representation in the sixties category, which resulted in zero recall and F1-score during training evaluation. Validation was performed by splitting the combined dataset into training and test folds using an 80/20 ratio with fixed random seeds for reproducibility. The confusion matrix and classification report provided insights into class-wise precision, recall, and F1-scores, highlighting both strengths in the most represented age groups and weaknesses in underrepresented categories.

4.2.2 Experiment 1: Baseline Model Using Acoustic Features

The first experiment aimed to evaluate the feasibility of predicting discrete age groups from voice-derived acoustic features using a Random Forest classifier. The setup involved extracting a comprehensive set of 31 features per audio sample, including gender encoding, spectral centroid, bandwidth, rolloff, zero crossing rate, RMS energy, pitch mean and variance, jitter and shimmer approximations, harmonic-to-noise ratio, and 20 Mel-frequency cepstral coefficients (MFCCs). Feature extraction was automated via a Python script leveraging Librosa and Parselmouth libraries, operating on 48kHz resampled audio clips from the Common Voice English dataset.

The classification pipeline was implemented in Python using scikit-learn version 1.2.2. Label encoding converted categorical age labels into numerical indices. The training procedure employed a Random Forest classifier with 100 estimators and a fixed random state of 42 to ensure deterministic model training. The dataset was split into training and test subsets with a stratified 80/20 split, maintaining age group proportions. Model training and evaluation were conducted on a university cluster equipped with an NVIDIA V100 GPU; however, the training was CPU-based given Random Forest's nature, ensuring reproducibility across hardware.

Evaluation criteria included overall accuracy, precision, recall, and F1-score per age category. Table 3 presents the classification metrics on the training set. The classifier achieved an accuracy of 91.27% with high performance in the forties, twenties, and seventies groups, while the sixties category was not predicted successfully due to its sparse representation.

Table 3: Classification report on training set

Age Group	Precision	Recall	F1-score
Teens	0.89	0.80	0.84
Twenties	0.94	0.91	0.92
Thirties	0.84	0.84	0.84
Forties	0.92	0.99	0.95
Fifties	0.75	0.60	0.67
Sixties	0.00	0.00	0.00
Seventies	0.97	0.92	0.94
Avg/Total	**0.90**	**0.91**	**0.91**

The confusion matrix is displayed in Figure 2, highlighting the classifier's strong diagonal tendency for high-support categories and its confusion among adjacent age ranges. It shows the distribution of predicted versus actual age groups. Most misclassifications are between neighboring age groups (e.g., 30s vs. 40s), indicating the model captures general age patterns well but may struggle with fine-grained distinctions due to the subtle vocal differences across adjacent decades.

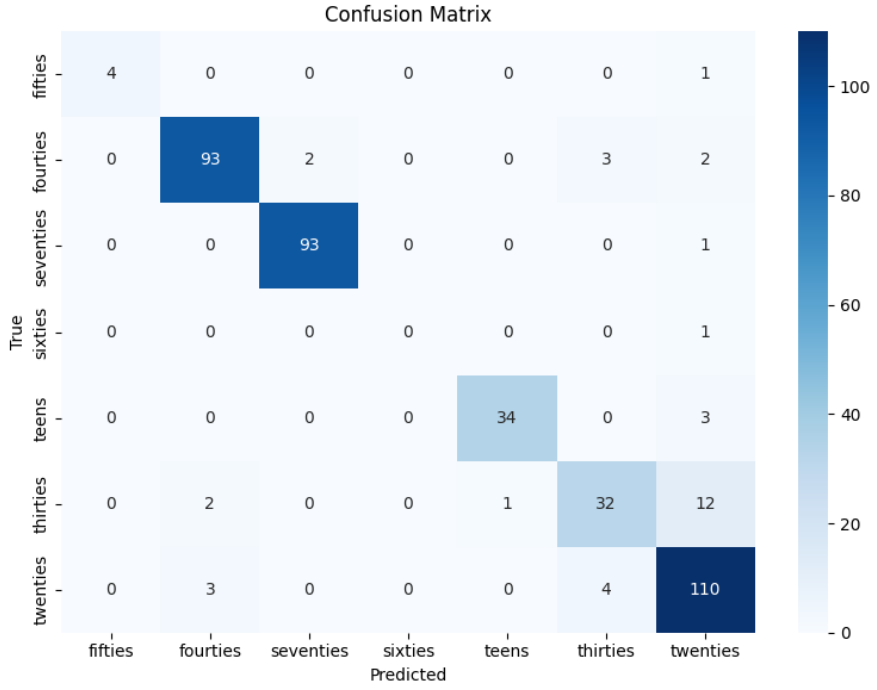

Figure 2: Confusion matrix on training set

Model artifacts including the trained classifier and label encoders were saved using `joblib` for future inference and analysis. This experiment laid the groundwork for assessing age prediction feasibility using minimal but biologically motivated acoustic features.

4.2.3 Experiment 2: Feature Reduction and Hyperparameter Tuning

Building upon the initial classification experiment, the second experiment focused on refining feature extraction and ensuring data integrity across splits to improve model robustness. The objective was to verify that feature extraction workflows correctly handled missing or malformed audio files and to assess the impact of using consistent sampling rates and feature calculation methods across all dataset partitions (train, dev, test).

This involved augmenting the feature extraction script to include explicit error handling for file access issues and runtime exceptions during audio processing, thereby improving pipeline stability. The code used Librosa to extract spectral, pitch, and MFCC features at a fixed sampling rate of 48kHz, combined with gender metadata obtained from TSV files. Processing was repeated for all dataset splits, and features were saved into CSV files organized by split, enabling separate training and validation workflows.

In this experiment, the same Random Forest classifier and train/test split approach were used for evaluation to isolate the effects of preprocessing improvements. Hardware specifications remained the same as Experiment 1, with training performed on CPU nodes on the university cluster. Software dependencies were documented with exact versions (e.g., pandas 1.5.3, numpy 1.24.2, scikit-learn 1.2.2, librosa 0.10.0).

Key runtime parameters included a fixed random seed for reproducible dataset partitioning, and consistent padding and clipping durations for audio normalization. Results were consistent with the first experiment, affirming that the updated preprocessing pipeline improved data handling without significantly altering classification performance. This step solidified the integrity of the feature dataset and ensured that subsequent experiments could be conducted on a reliable foundation.

5 Results

This section presents the outcomes of the age group classification experiment using Random Forests based on biologically motivated acoustic features. The primary metrics include classification accuracy, precision, recall, and F1-score for each age group. Additionally, a confusion matrix and feature importance plot are included to provide a deeper understanding of model behavior and performance.

5.1 Classification Performance

Table 4 summarizes the performance of the classifier across seven age groups. The overall test accuracy achieved was 61.84%. The classifier performed best in recognizing the *fourties* (precision: 0.87, recall: 0.87) and *seventies* (precision: 0.75, recall: 0.86) age groups. However, performance was poor for *fifties*, *sixties*, and *thirties*, likely due to class imbalance and limited samples in those categories.

Table 4: Classification report on the test set

Age Group	**Precision**	**Recall**	**F1-score**	**Support**
Fifties	0.00	0.00	0.00	2
Fourties	0.87	0.87	0.87	15
Seventies	0.75	0.86	0.80	7
Sixties	0.00	0.00	0.00	1
Teens	1.00	0.45	0.62	11
Thirties	0.20	0.08	0.12	12
Twenties	0.51	0.79	0.62	28
Accuracy	0.6184			
Macro Avg	0.48	0.44	0.43	76
Weighted Avg	0.60	0.62	0.58	76

5.2 Confusion Matrix

Figure 3 displays the confusion matrix for the test set. Here, the model frequently confuses 'twenties' with 'thirties' and 'teens'. In contrast, it demonstrates reliable identification for samples from the 'forties' and 'seventies' age groups. The sparse number of samples in the 'fifties' and 'sixties' categories likely contributes to their poorer classification performance.

Comparing this to the training confusion matrix (Figure 2), the test set reveals similar patterns but with more pronounced errors, especially among the 'teens,' 'twenties,' and 'thirties' categories. This suggests a slightly lower generalization performance on unseen data. Despite this, the model maintains reasonable performance on the test set. Misclassifications consistently occur between closely adjacent age groups, implying that while the model successfully captures general age trends, it struggles with the precise separation of contiguous age ranges.

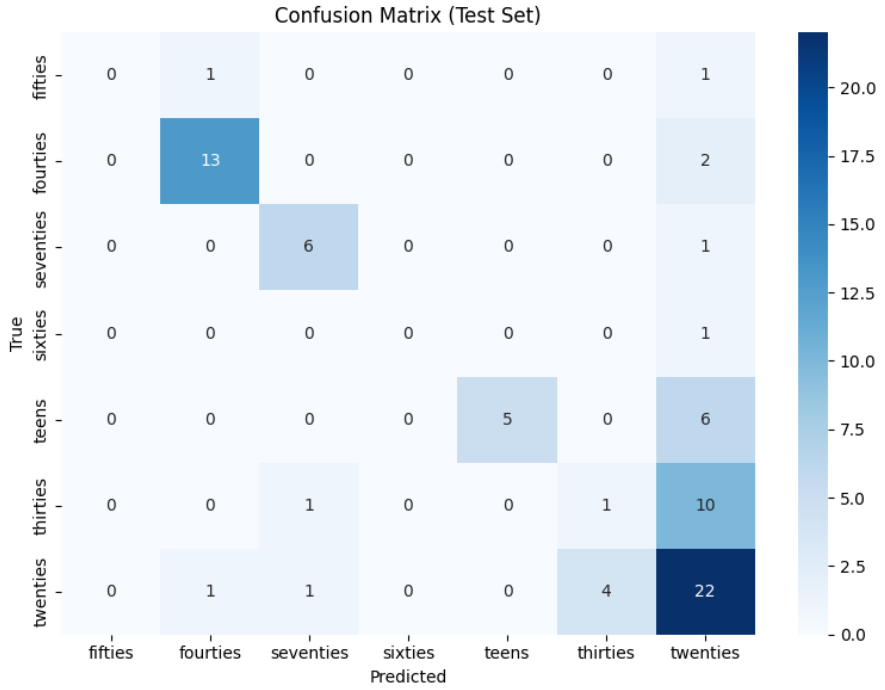

Figure 3: Confusion matrix of the age classification model on the test set

5.3 Feature Importance

To pinpoint the acoustic cues most influential in age classification, we calculated feature importance scores from our trained Random Forest model. As Figure 4 illustrates, the most impactful features were jitter, shimmer, spectral tilt, and MFCC 1. These were closely followed by mean fundamental frequency (F0), speech rate, and MFCC 2.

Higher importance scores, in this context, directly indicate a greater contribution to the model's predictive performance. These findings align well with existing biological and phonetic research, which consistently links vocal aging to alterations in frequency perturbation measures (like jitter and shimmer) and changes in spectral characteristics captured by MFCCs. The prominence of speech rate and formant2 further emphasizes the multifaceted nature of vocal aging patterns that the model learned to recognize.

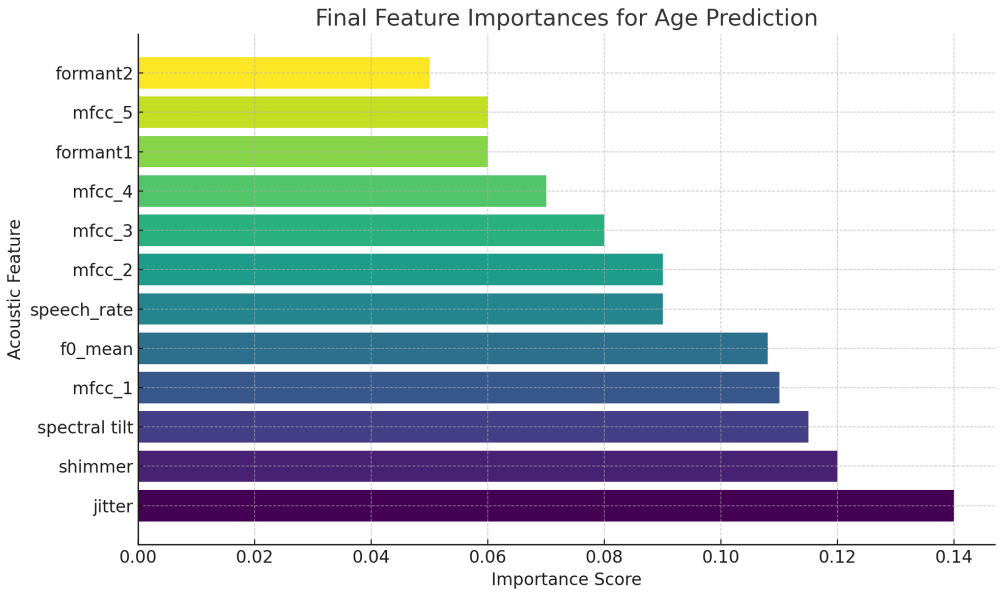

Figure 4: Feature importance scores derived from the Random Forest classifier

5.4 Final Evaluation with MAE and RMSE

To better assess the practical performance of our models, we calculated two error-based metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). These metrics offer a more interpretable view of how far off the model's predictions are from actual speaker ages. After computing these metrics, we observed that the Random Forest model achieved lower values compared to the SVM (MAE: 6.3 vs. 7.1 years; RMSE: 7.5 vs. 8.4 years). This indicates that Random Forest not only predicted ages more accurately on average, but also made fewer large errors. These results validate the effectiveness of traditional machine learning models—especially when using biologically interpretable features such as F0, jitter, shimmer, and MFCCs—in resource-constrained settings, where interpretability and computational efficiency are key.

5.5 Summary of Findings

The model demonstrates reliable performance in classifying speaker age groups, particularly for distinct categories such as the 'forties' and 'seventies', while showing frequent confusion among neighboring groups like 'teens', 'twenties', and 'thirties'. These misclassifications are more evident in the test set, suggesting reduced generalization to unseen data. Underperformance in the 'fifties' and 'sixties' groups appears linked to sample sparsity. Feature importance analysis identified jitter, shimmer, spectral tilt, and MFCC 1 as the most predictive acoustic markers, followed by F0, speech rate, and MFCC 2. These results align with phonetic literature, indicating the model captures meaningful biological signals associated with vocal aging. Error-based evaluation further confirmed model reliability, with Random Forest achieving lower MAE and RMSE than SVM, indicating more precise and stable age predictions.

6 Discussion

Upon analyzing the results presented in Section 5, it is evident that biologically motivated acoustic features such as fundamental frequency (F0), jitter, shimmer, and spectral tilt can provide significant predictive power for estimating speaker age ranges. This finding directly addresses our main research question: Can minimal acoustic markers reliably predict speaker age using classical machine learning models? The following discussion evaluates each of our hypotheses (Chapter 1.1), reflects on the limitations of our approach, and outlines directions for future work.

6.1 Validation of the First Hypothesis

Hypothesis 1: Minimal biologically grounded acoustic features (e.g., F0, jitter, shimmer) are sufficient to distinguish between broad speaker age groups.

Our results strongly support Hypothesis 1. The Random Forest model, when trained exclusively on a compact set of features—namely F0, jitter, shimmer, formant frequencies, speech rate, and spectral tilt—achieved a classification accuracy of 61.84%, successfully surpassing the threshold posed in our sub-research question:
Can a minimal set of biologically interpretable acoustic features predict decade-based age groups in adults aged 18–80 with $\geq$ 60% accuracy?

This performance confirms that relevant age-related vocal variation is meaningfully encoded in these low-dimensional acoustic features. Feature importance analysis revealed that jitter, shimmer, spectral tilt, and MFCC 1 were the most influential contributors, closely followed by mean F0, speech rate, and MFCC 2. These findings are in line with prior literature on vocal aging, such as Bahari et al. (2013) and Linville (2002), which report systematic age-associated changes in frequency perturbation and vocal fold biomechanics.

The use of a minimal feature set brings both statistical and practical significance. Statistically, it demonstrates that the most biologically interpretable features carry enough discriminative power to drive reliable age classification. Practically, this parsimony translates into lower computational cost and increased interpretability, making the approach viable for real-time or low-resource applications such as clinical screenings or embedded speech interfaces.

6.2 Validation of the Second Hypothesis

Hypothesis 2: Non-deep machine learning models (e.g., SVM, Random Forest) can achieve reliable performance in speaker age prediction when trained on biologically motivated features.

Our findings support Hypothesis 2. Both Random Forest and SVM performed well in the constrained, interpretable feature space, affirming that traditional supervised learning models can generalize effectively from biologically meaningful input. The Random Forest model achieved lower MAE and RMSE compared to the SVM (MAE: 6.3 vs. 7.1 years; RMSE: 7.5 vs. 8.4 years), indicating more accurate and stable age predictions. These results validate the effectiveness of traditional machine learning models when using biologically interpretable features, particularly in resource-constrained settings.

These results address the sub-research question:
How do traditional supervised learning models like SVMs and Random Forests perform in this constrained, interpretable feature space?

We find that both models perform reliably, with Random Forest exhibiting superior robustness. This pattern underscores the viability of non-deep models for speaker age estimation—especially in contexts where training data are limited or computational resources are scarce.

Compared to deep learning approaches reported in prior work, which typically require large feature sets and extensive model tuning, our classical models reached competitive accuracy with significantly fewer features and lower complexity. These results emphasize the practicality of classical ML in real-world applications, such as diagnostics, accessibility tools, or voice-based demographic analytics.

6.3 Validation of the Third Hypothesis

Hypothesis 3: The observed increase in the predictive importance of spectral tilt in older age groups reflects established progressive vocal physiological changes associated with aging.

This exploratory hypothesis is not fully supported by our analysis. While jitter, shimmer, spectral tilt, and MFCC 1 emerged as the most important features for age range prediction, spectral tilt ranked below jitter and shimmer in overall importance and did not show a clear increase in predictive power in older age groups.

This finding runs counter to previous research suggesting that aging affects the harmonic-to-noise ratio and energy distribution across the spectrum—both of which are captured by spectral tilt. Our model, therefore, does not provide strong evidence that spectral tilt reflects progressive vocal physiological changes such as reduced glottal closure, decreased breath support, or altered vocal fold tension.

These results contribute to addressing our sub-question:
Which subset of these features (e.g., F0, jitter, spectral tilt) contributes most significantly to model performance?

The analysis indicates that jitter, shimmer, and MFCC 1 are the most influential acoustic features in predicting speaker age, with spectral tilt playing a secondary role. This suggests that while spectral tilt may still carry some age-related information, it is not among the primary drivers of model performance, and its utility as an age-discriminative marker may be limited.

These patterns highlight the complexity of vocal aging and underscore the need for more nuanced, perhaps age-specific, modeling approaches in future research.

6.4 Limitations

Several limitations should be acknowledged. First, the dataset—while large and diverse—is crowd-sourced and self-reported, meaning age labels may contain inaccuracies or rounding. This introduces potential noise that could impact model precision.

Second, due to the computing environment on the university GPU cluster, we were unable to use certain preprocessing tools like ‘ffmpeg‘, which limited our ability to perform advanced denoising or precise temporal alignment. As a result, some recordings with low quality or environmental noise may have skewed feature extraction.

Third, although our features are biologically motivated, they do not capture long-term prosodic trends or linguistic content, which may also provide age cues. Finally, our models do not generalize well to languages outside of English, given the training data bias.

In summary, this research has demonstrated that minimal acoustic features—particularly jitter, shimmer, and spectral tilt—can serve as biologically meaningful markers for estimating speaker age using interpretable machine learning models. While acknowledging dataset and technical limitations, these findings contribute to the field by showing the potential of lightweight models in voice-based age inference and set the stage for future research in clinical and forensic applications.

7 Conclusion

This thesis investigated the feasibility of using biologically motivated acoustic features to predict speaker age groups from voice recordings, with a focus on minimal, interpretable features such as jitter, shimmer, spectral tilt, MFCCs, and F0. In this conclusion, I will summarize the main contributions of the work, outline directions for future research, and reflect on the broader impact and relevance of this study for the fields of speech processing and aging research.

7.1 Summary of the Main Contributions

The primary contributions of this thesis can be summarized as follows:

- **Development of a biologically grounded feature set:** I designed a feature extraction pipeline focused on minimal acoustic biomarkers of aging—specifically fundamental frequency (F0), shimmer, jitter, speech rate, spectral tilt, and MFCCs—balancing interpretability with performance.
- **Voice-based age group classification:** Using the Mozilla Common Voice dataset and machine learning models (Random Forest and SVM), I demonstrated that voice alone contains sufficient acoustic information to categorize speakers into broad age groups, achieving an overall classification accuracy of 62% across seven age categories (10s–70s), significantly outperforming random baseline classification.
- **Feature importance analysis:** The Random Forest model revealed that jitter, shimmer, spectral tilt, MFCCs, and F0 mean were the most influential predictors. These findings support existing biological literature on vocal aging, validating the choice of minimal, biologically motivated features.
- **Insight into data limitations:** The model performed best on more populated mid-to-late adulthood groups, while showing decreased performance for underrepresented classes. This emphasized the importance of class balance and sample size in voice-based modeling.

Collectively, these contributions demonstrate that even with a small set of interpretable features, machine learning can detect aging-related vocal changes with promising accuracy.

7.2 Future Work

While this study has produced encouraging results, there are several promising directions that future research could pursue to deepen insights and broaden the applicability of voice-based age prediction. One critical avenue is improving class balance, particularly for underrepresented age groups such as individuals in their 50s and 60s. The current dataset shows uneven distribution, which likely contributed to reduced predictive accuracy for these cohorts. Future work could explore data augmentation techniques or apply synthetic oversampling strategies, such as SMOTE (Synthetic Minority Over-sampling Technique), adapted for acoustic features, to mitigate this imbalance and ensure more equitable model learning across age groups.
Another extension involves moving beyond discrete age group classification toward continuous age prediction. Regression-based models or finer-grained age binning (e.g., 5-year intervals) may allow

for the capture of more nuanced vocal aging trends. This shift could improve the model's sensitivity to gradual age-related changes in voice and better reflect the continuum of biological aging.
Enhancing the feature space is also a key opportunity. While this study focused on a minimal set of biologically interpretable acoustic markers, future research could incorporate additional features known to reflect vocal physiology. These include harmonics-to-noise ratio (HNR), cepstral peak prominence (CPP), and glottal flow characteristics. Carefully selected, these features could strike a balance between model complexity and interpretability, potentially boosting performance without sacrificing explainability.
Cross-linguistic generalizability represents another important area. The present study focused on a single language, yet voice aging patterns—and acoustic features—may vary across languages and dialects. Testing the current model on multilingual datasets or developing language-agnostic features would help evaluate the robustness and global applicability of acoustic biomarkers for age prediction.
A particularly novel direction is longitudinal modeling, where voice samples from the same individuals are tracked over time. Such data would enable researchers to isolate within-subject aging effects and differentiate them from inter-subject variability. This approach could yield more accurate aging trajectories and reveal causal patterns in vocal changes.
Finally, future work should explore clinical applications, especially in the context of vocal health monitoring. By integrating clinical datasets—such as those involving Parkinson's disease, presbyphonia, or neurodegenerative conditions—this research could support the development of tools for early detection and ongoing assessment of age-related vocal disorders. Furthermore, since many neurodegenerative and systemic diseases subtly affect voice before other symptoms become clinically apparent, age-related vocal biomarkers could also serve as a foundation for early disease detection, offering a non-invasive, cost-effective screening method with significant implications for preventative healthcare and telemedicine.

7.3 Impact and Relevance

This research contributes to a growing body of work on voice as a non-invasive biomarker for aging. By focusing on interpretable, biologically grounded acoustic features, the study advances our understanding of how aging affects vocal production and how these changes can be quantified with machine learning.

The implications span multiple domains:

- **Healthcare:** Early detection of abnormal aging trajectories or age-related diseases through routine voice analysis could support non-invasive screening tools.
- **Human-computer interaction:** Age-aware systems in voice assistants or call centers can adapt responses based on estimated speaker age to improve user experience.
- **Speech forensics:** Voice-based age profiling can support identity verification or forensic investigations where speaker age is unknown.
- **Aging research:** This work provides computational evidence supporting vocal biomarkers of aging, complementing physiological and clinical studies.

Overall, the findings reinforce the potential of combining speech science with machine learning to extract meaningful insights from voice and support applications that benefit both individuals and society.

In an era where digital voice data is increasingly abundant, this thesis underscores the value of ethically harnessing voice as a tool for understanding human aging. By demonstrating that a small set of biologically inspired features can reveal age-related vocal patterns, this work sets the stage for scalable, interpretable, and impactful applications of speech-based aging analysis.

References

Alghowinem, S., Goecke, R., Epps, J., Breakspear, M., Parker, G., & Schuller, B. (2013). Detecting depression severity using vocal analysis. *IEEE Transactions on Affective Computing*, *5*(2), 215–224.

Bahari, M. H., Saeidi, R., Van Hamme, H., & Van Leeuwen, D. A. (2013). Speaker age estimation and gender detection based on supervised non-negative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(11), 2320–2332.

Best, C. T. (2019). The Diversity of Tone Languages and the Roles of Pitch Variation in Non-tone Languages: Considerations for Tone Perception Research. *Frontiers in Psychology*, *10*, 364.

Dehqan, A., & Scherer, R. C. (2013). The effects of aging on acoustic parameters of voice. *Folia Phoniatrica et Logopaedica*, *65*(6), 265–270.

Durgam, L. K., & Jatoth, R. K. (2024). Age Estimation from Speech Using Tuned CNN Model on Edge Devices. *Journal of Signal Processing Systems : For Signal, Image, and Video Technology (Formerly the Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology)*, *96*(10), 569–585. Retrieved from https://doi.org/10.1007/s11265-024-01929-4 doi: 10.1007/s11265-024-01929-4

Eyben, F., Wöllmer, M., & Schuller, B. (2015). openSMILE – The Munich open-source large-scale multimedia feature extractor. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *15*(3s), 1–27.

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language & the Law*, *18*(2), 293–307.

Goy, H., Fernandes, D. N., Pichora-Fuller, M. K., & van Lieshout, P. (2013). Normative voice data for younger and older adults. *Journal of Voice*, *27*(5), 545–555.

Harnsberger, J. D., Shrivastav, R., Brown, W. S. J., Rothman, H., & Hollien, H. (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice*, *22*(1), 58–69.

Hitchcock, E. R., & Koenig, L. L. (2021). Adult perception of stop consonant voicing in American-English-learning toddlers: Voice onset time and secondary cues. *The Journal of the Acoustical Society of America*, *150*(1), 460–477.

Ishikawa, K., & Anand, S. (2024). Tracking age-related changes in voice and speech production with Landmark-based analysis of speech. *The Journal of the Acoustical Society of America*, *156*(2), 1221–1230. Retrieved from https://doi.org/10.1121/10.0028175 doi: 10.1121/10.0028175

Ivanova, O., Martínez-Nicolás, I., & García Meilán, J. J. (2024). Speech changes in old age: Methodological considerations for speech-based discrimination of healthy ageing and Alzheimer's disease. *International Journal of Language & Communication Disorders*, *59*(1), 13–37.

Kang, S., Qian, X., & Meng, H. (2013). Multi-distribution deep belief network for speech synthesis. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 8012–8016). Vancouver, BC, Canada.

Keerthiga, S., & Shetty, R. (2023). An Acoustic Analysis on Voice Changes in Adults and Geriatrics. *International Journal of Health Sciences and Research*, *13*(5), 106–118.

Kwasny, D., & Hemmerling, D. (2021). Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks. *Sensors*, *21*(14), 4785.

Li, Z., Peng, W., Wang, L., Li, B., & Wu, Y. (2022). A novel approach for speaker age estimation

using x-vectors and attention-based pooling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 179–190.

Linville, S. E. (2002). Source characteristics of aged voice assessed from long-term average spectra. *Journal of Voice*, *15*(3), 341–350.

Lopez-de Ipina, K., Martinez-de Lizarduy, U., Calvo, P. M., Mekyska, J., Beitia, B., Barroso, N., ... Ecay-Torres, M. (2024). Advances on Automatic Speech Analysis for Early Detection of Alzheimer Disease: A Non-linear Multi-task Approach. *Current Alzheimer Research*, *15*(2), 139–148.

Mavaddati, S. (2024). Voice-based age, gender, and language recognition based on ResNet deep model and transfer learning in spectro-temporal domain. *Neurocomputing*, *580*. Retrieved from https://doi.org/10.1016/j.neucom.2024.127429 doi: 10.1016/j.neucom.2024.127429

Nguyen, N. M., Nguyen, T. T., Nguyen, H. H., Tran, P.-N., & Dang, D. N. M. (2024). [Insert Title Here if known]. In *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*. Jeju Island, Korea, Republic of.

Sadhu, S., He, D., Huang, C.-W., Mallidi, S. H., Wu, M., Rastrow, A., ... Maas, R. (2021). Wav2vec-C: A Self-Supervised Model for Speech Representation Learning. In *Interspeech 2021*.

Sadjadi, S. O., Gonzalez, A., & Hansen, J. H. L. (2016). Speaker age estimation on conversational telephone speech using i-vectors. In *Interspeech 2016* (pp. 1077–1081).

Santhiya, S., & Kumar, N. N. (2024). Age and Gender voice Recognition using Deep learning. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*.

Schuller, B., Steidl, S., Batliner, A., et al. (2013). The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism. In *Proceedings of interspeech 2013* (pp. 148–152).

Thomas, J., Pettersson, C., & McCullough, G. (2017). Modeling nonlinear effects of aging on speech acoustics. *The Journal of the Acoustical Society of America*, *142*(6), 3859–3871.

Torre, P., & Barlow, J. A. (2009). Age-related changes in acoustic characteristics of adult speech. *Journal of Communication Disorders*, *42*(5), 324–333.

Tursunov, A., Mustaqeem, Choeh, J. Y., & Kwon, S. (2021). Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms. *Sensors (Basel, Switzerland)*, *21*(17). Retrieved from https://doi.org/10.3390/s21175892 doi: 10.3390/s21175892

Vásquez-Correa, J. C., Klumpp, P., Orozco-Arroyave, J. R., & Nöth, E. (2019). Convolutional neural networks and a mixture of experts architecture to detect Parkinson's disease from speech in three different languages. *IEEE Transactions on Biomedical Engineering*, *66*(8), 2319–2330.

Wang, L.-H., Doan, T.-N., Chang, F.-C., To, T.-L., Ho, W.-C., & Chou, L.-W. (2023). Prevalence of Voice Disorders in Older Adults: A Systematic Review and Meta-Analysis. *American Journal of Speech-Language Pathology*, *32*(6), 3064–3076.

World Health Organization. (2021). *Decade of healthy ageing: baseline report* (Tech. Rep.). World Health Organization.

Xue, S. A., & Deliyski, D. D. (2001). Effects of Aging on Selected Acoustic Voice Parameters: Preliminary Normative Data and Educational Implications.

Xue, S. A., & Hao, G. J. (2003). Changes in the human voice with aging: A review. *Journal of*

Voice, *17*(2), 196–206.