U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Impact of Vocal Traits Distribution on Speech Applications' Performance and Bias

André Luis Monforte Neves Azenha de Almeida

Dissertation

Master in Modelling, Data Analysis and Decision Support Systems

Supervised by

Professor Doutor João Gama

Doutor Rui Correia

2021

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Professor Doutor João Gama, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would like to acknowledge my colleagues from my internship at DefinedCrowd Corp. for their wonderful collaboration. I would particularly like to single out my supervisor, Doutor Rui Correia. Rui, thank you for your patient support and for all of the opportunities I was given to further my research.

In addition, my gratitude also extends to my family and friends who have been assisting, supporting and caring for me all of my life.

Abstract

As algorithms drive more decision-making processes, machine learning models' tendency to learn our input data biases is a massive problem. Furthermore, the wide range of new diverse and heterogeneous users demands robust and unbiased solutions that perform successfully regardless of their individual characteristics or demographics. More than ever, companies are starting to be held accountable for their models' behaviors and performance, especially concerning minorities and marginalized groups.

Research has been conducted concerning possible bias conflicts in speech applications, having identified systematic errors against social groups, such as females, elderlies, and misrepresented ethnicities. To fight this, data providers' prevalent interventions focus on assuring uniform distributions over binary gender groups. Such interventions, however, have three major limitations. First, gender proxies are non-descriptive of the vocal characteristics they are trying to emulate, hence may not represent the complete spectrum of diversity. Secondly, in situations where the collection is not done in person (namely via crowdsourcing platforms), they are easy to mask if the contributors are ill-intentioned and difficult to validate from the requester's point of view. Finally, if misused, these proxies may be perpetuating social stereotypes (eg., what a male voice is expected to sound like).

This work explores the hypothesis of replacing gender proxies with actual vocal representations of the speaker to drive the data collection process. Models would be solely based on domain-specific (ethical) information and balanced over particularities of the speakers' voice (for instance, pitch, and volume) instead of proxies to the desired descriptors.

Results show that, when compared to the prevalent method based on self-reported gender labels, vocal traits (particularly pitch and spectral centroid) offer a more verifiable, effective and ethical approach to the speech data collection: verifiable since they are measurable and objective depictions of the speaker; effective since they improve performance by two percentage points and reduce bias both across gender and age groups; and ethical in the sense that they are actual and fact-based representations, blind to the speaker´s ethnicity, age, gender, etc.

Keywords: AI Fairness; Input Bias; Voice Profiling; Speech Applications; Speech Recognition

Resumo

A utilização crescente de algoritmos de *Machine Learning* em processos fundamentais da nossa sociedade acarreta o risco de estes sistemas replicarem e amplificarem padrões de discriminação vigentes na nossa sociedade. Além disso, a nova e heterogénea gama de utilizadores exige soluções robustas e imparciais que funcionem com sucesso, independentemente de características demográficas ou individuais do indivíduo. Com efeito, diversas empresas e instituições a nível mundial têm vindo a ser responsabilizadas pelo comportamento e desempenho dos seus modelos, especialmente no que diz respeito a minorias e grupos marginalizados.

No caso específico de aplicações de reconhecimento e processamento de voz, têm sido identificados erros sistemáticos contra diversos grupos da nossa sociedade, como mulheres, idosos, ou etnias minoritárias. Como resposta a estes problemas, os provedores de dados têm-se focado em garantir uma representação uniforme de indivíduos nos dados de treino, utilizando como critério predominante o género (binário) do indivíduo. Tais intervenções têm, no entanto, três grandes limitações. Primeiramente, o género não é uma descrição objetiva da voz, sendo um mero *proxy* da voz do indivíduo. Consequentemente, tais *proxies* podem representar errada ou insuficientemente o espectro de diversidade dos indivíduos. Em segundo lugar, em situações em que a recolha de dados não é realizada pessoalmente (nomeadamente através de plataformas de *crowdsourcing*), as informações de género são facilmente falseadas por utilizadores mal-intencionados, e difíceis de validar. Finalmente, se usados indevidamente, estes *proxies* podem perpetuar estereótipos sociais (por exemplo, como se espera que uma voz masculina soe).

Assim, esta dissertação explora a hipótese de usar representações objetivas da voz para garantir uma representação uniforme de indivíduos nos dados de treino. Os modelos treinados segundo este método seriam balanceados segundo particularidades da voz (por exemplo, tom e volume da voz), ao invés de *proxies* como o género do indivíduo.

A análise dos dados revelou que uma representação uniforme das características da voz nos dados de treino oferece uma abordagem mais verificável, eficaz e ética à recolha de dados de fala em comparação com os proxies de género. Verificáveis uma vez que são representações mensuráveis e objetivas do orador. Eficazes uma vez que melhoram o desempenho em dois pontos percentuais e aumentam a imparcialidade do modelo entre grupos de género e etários. Ética no sentido em que é baseada em representações reais e factuais do indivíduo, i.e., independentes da etnia, idade e género do orador.

Palavras-Chave: IA Responsável; Discriminação; Aplicações de Reconhecimento e Processamento de Voz; Reconhecimento de Fala

List of Siglas, Abbreviations and Acronyms

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
CTC	Connectivist Temporal Classification
CV	Coefficient of Variation
DNN	Deep Neural Network
F0	Fundamental Frequency
FC	Fully-Connected Network
GMM	Gaussian-Markov Model
GTCC	Gamma-Tone Cepstral Coefficients
HIT	Human-Intelligence Tasks
HMM	Hidden Markov Models
HNR	Harmonic-to-Noise Ratio
HTK	Hidden Markov Model Tool Kit
IQR	Interquartile Range
LPC	Linear Predictive Coding
LPCC	Linear Prediction Cepstral Coefficients
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
NLP	Natural Language Processing
PLP	Perceptual Linear Predictive
pp	Percentage points
RASTA-PLP	Relative Spectral Transform PLP
ReLU	Rectified Linear Activation Function
RMS	Root Mean Square Energy
RNN	Recurrent Neural Network
STFT	Short-Time Fourier Transformation
TDNN	Time-Delay Neural Network
UBM	Universal Background Model
USA	United States of America
VUI	Voice User Interfaces
WER	Word Error Rate
WSFT	Weighted Finite-State Transducers

Contents

List of Figures

List of Tables

1 | Introduction

1.1 Motivation

One of the most prominent discussions of the XXI century is the fight against discrimination. Whether it is racial, gender, or another type of discrimination, human societies have inherent unfairness conflicts in their core foundations. Data, as a mirrored representation of reality, is no different. Indeed, one of the current central issues in Artificial Intelligence (hereinafter, AI) is fairness.

Broadly, fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making [1]. Because machines can treat similarly-situated people and objects differently, research is starting to reveal some troubling examples in which the reality of algorithmic decision-making falls short of our expectations. Some algorithms run the risk of replicating and even amplifying human biases, particularly those affecting protected groups [2]. Prevailing examples come from facial recognition systems biased against black-skinned users[1], or even hiring algorithms that systematically downgraded the score given to women's *resumés*[2]. Naturally, given this succession of conflicts, companies are starting to be held accountable for their models' behaviors and performance, creating new constraints and needs for AI.

In the context of AI, fairness entangles two major premises. First, a fair system should only act based on domain-specific (and ethical) information; and, secondly, its performance should be comparable across distinct classes of its user base. Certainly, the violation of one of the referred premises leads to unfair systems, which may be the result of poorly developed AI. The causes of these conflicts can be drilled-down to the key components of AI systems: inappropriate methodology (algorithms, features, testing, monitoring) and biased data. The latter, *data* or *input bias*, is the focus of our research.

Notably, an AI system's quality depends significantly on the volume and quality of the data used in its training. The latter leaves us facing the emerging concern that "AI artifacts tend to reflect the goals, knowledge, and experience of their creators" [3]. Indeed, if historical biases are factored into our training sets, the existing prejudices will be captured by the models, and potentially reproduced and amplified. Besides historical issues, input bias can emanate from incomplete or unrepresentative data. If an algorithm is more representative of some people than others, the model may systematically go against unrepresented or under-representative groups.

Subsequently, input bias is potentially manifested in all forms of AI, namely in systems that are

[1] https://www.wired.com/story/can-apples-iphone-x-beat-facial-recognitions-bias-problem/

[2] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

now part of our everyday lives, such as Voice-user interfaces (VUI). VUI´s, like *SIRI* and *Alexa* [4], make spoken human interaction with computers possible, using speech recognition to understand spoken commands and answer questions, and typically text to speech to play a reply. VUI's have been added recently into automobiles, home automation systems, or even virtual smartphones' assistants. Naturally, this widespread adoption of speech as an interface led to a growing demand for data to train these systems.

As data is said to be the *new oil*[3], few data sets are publicly available for speech systems [5]. The best known are corpora like TIMIT [6], or Switchboard [7], which date back to the early 1990s. Such scarceness is mostly related to how costly gathering and annotating audio data with respect to money and time. This led to the increasing importance of data providers, such as crowdsourcing platforms. Crowdsourcing is a particularly attractive solution for large data collection efforts, allowing to reach a diverse crowd in a expedite and scalable manner (both faster and less costly). Briefly, it consists of gathering and distributing work across a large pool of human contributors, typically via an online platform. Once completed, payment is given to the contributors, proportional to their participation.

Crowdsourcing, however, presents a new set of challenges, including the loss of control by data requesters and the vulnerability to ill-intentioned contributors [8]. To address these issues, quality ensuring is often done by submitting these recordings to validation tasks. Successfully validated recordings are then packaged for delivery, along with any additional relevant metadata (with respect to the recordings – such as the text that was recorded – or the speakers – such as gender or age).

However, different aspects of the data present different validation challenges. For instance, validating that a recording has no background noise is more straightforward than validating that the speaker is of a specific gender. Indeed, self-reported speaker information (provided when signing up for the platform, such as gender or age) is sensitive, and ultimately hard to verify and contest.

Gender stats have, indeed, an important role in the speech collection pipeline: they are used as a relevant tool to mitigate the recently unveiled bias conflicts in speech applications. Garnerin et al. [5] identified bias towards women in the performance of speech recognition systems by analyzing the gender representation in 4 major corpora of French broadcast. The authors concluded that the disparity of available data for both genders caused performance to decrease on women. Vipperla [9] identified that the word error rates (WER) of Automatic Speech Recognition systems (ASR) are significantly higher for older adults, when compared to younger adults. Bias was also identified against racial groups. According to Koenecke et al. [10], black speakers have an average WER 10 p.p. greater than white speakers when using state-of-the-art speech recognition systems — developed by Amazon, Apple, Google, and Microsoft.

[3]https://www.theguardian.com/technology/2013/aug/23/tech-giants-data

1.2 Problem Discussion

As an answer to the various bias conflicts unveiled in speech applications, data providers have been focusing on balancing speech datasets over binary gender groups. Such interventions, however, pose three major limitations. First and foremost, gender is only a proxy for actual vocal characteristics of the speaker, and for that reason, may not represent the complete spectrum of speaker diversity. Secondly, as previously referred, these interventions are based on self-reported data, which is hard to contest. Finally, limiting speaker classification to gender labels perpetuates social stereotypes, for instance, of what a male voice is expected to sound like.

Given this background, this work explores the hypothesis of replacing gender proxies with actual vocal representations of the speaker to drive the data collection process. The identified vocal traits could then be used as criteria for balancing the training sets for speech applications. It is important to note that measuring systems' performance across social groups (like the ones provided by gender information) is still relevant. However, this kind of sensitive self-reported metadata must not be contested on the basis of normative (and potentially offensive) approaches, and for that reason, are not fit to drive data collection. Accordingly, our hypothesis is that the identified vocal traits offer a more verifiable and ethical way to describe speech data, ultimately improving performance and reducing bias.

To test this hypothesis, we will evaluate the impact of vocal traits in speech applications' performance using a concrete speech application (automatic speech recognizer). *Post-hoc* analysis on the systems´ performance should allow us to conclude on which vocal traits should be uniformly represented in the training dataset and measuring the impacts of such distribution in the model's performance and biases. Our research was guided by the following two questions:

1. Which voice traits better differentiate and characterize speakers?
2. What is the impact of balancing such features in the training dataset of a speech application?

This document is structured around six different sections. Chapter 2 contains a literature review that describes relevant work on voice profiling, speech recognition systems, crowdsourcing data collection, and bias. Chapter 3 explains the proposed methodology. Chapters 4 and 5 detail the obtained results for each of our research questions. Finally, Chapter 6 discusses the obtained results and evaluates our initial hypothesis, and Chapter 7 presents the conclusions of our study and makes some remarks on future work.

2 | Literature Review

This work sits at the intersection of three research topics with a wide array of work: crowdsourcing, speaker profiling and AI fairness. For this purpose, this chapter is structured as follows: Section 2.1 sets some terminology on the phonetics of speech sound; Section 2.2 reviews state-of-the-art voice profiling techniques with a specific focus on gender categorization through voice; Section 2.3 provides some background on the high-level architecture for automatic speech recognition systems, and Section 2.4 introduces terminology and background work on crowdsourcing data collection. Finally, Section 2.5 reviews the key bias and fairness concepts in AI, with a specific focus on bias conflicts detected in speech applications.

2.1 Phonetics

2.1.1 Phonemes and Phones

Humans produce, classify, and interpret audio signals all the time without conscious effort. Indeed, Humans are inherently set to capture hearable sounds[1], and to decode the information carried by the signal. In this work, our focus will rest on a specific category of hearable sound: speech sound.

Speech consists of sequences of sounds, mostly continuous sounds, both within words and across word boundaries. Indeed, a speaker can quickly dissect their constant sounds into words and split words into component sounds. For example, three components can be distinguished in the word *bat*, corresponding to the letters *b*, *a*, and *t*. The distinct units of sound in spoken English distinguish one word from another. For example, when we switch the /*b*/ in *bat* with /*k*/, we reproduce another word, *cat*.

The sound segments above are **phonemes**, the smallest sound unit that distinguishes meaning between sounds in a given language [12]. **Phones** are its acoustic realization. Concisely, phonemes are an abstract concept in linguistics to distinguish words, and phones are how we pronounce them. English as spoken in the United States contains a total of 44 phonemes[2]. Finally, **allophones** are the context-dependent (CD) representation of phones, i.e., comprise the pronunciation variations of a given phone. These are particularly useful to capture accent and pronunciation variations for a given phone. Identifying allophones in captured sounds –, i.e, understanding how humans perceive and produce speech –, is the initial task of speech recognition systems.

[1] The Gerhard [11] taxonomy separates hearable sounds into Noise Natural, Artificial, Speech, and Music.

[2] Most languages have between 20 to 60 phones in their vocabulary.

2.1.2 Production of Phonemes

Speech sounds are generated whenever we tense up our vocal folds, creating the *glottal pulse*. When we exhale air from the lungs, it pushes the vocal folds open. The airstream generates a vibration of the vocal folds, producing a sound wave. These open and close cycles create a series of sound wave frequencies with a fundamental frequency that tends to differ between male and female speakers. Indeed, Singh [13] states that the adult woman's average pitch range is from 165 to 255 Hz, while a man's is 85 to 155 Hz.

Next, the *vocal articulators* (such as teeth, nasal cavity, and tongue) create different vocal tract shapes, generating different resonances. They act as filters in suppressing or amplifying output sound frequencies. The final generated sound can thus be seen as a weighted additive combination of different frequencies, which literature refers to as *frequency components* [13]. The scheme below summarizes the described process.

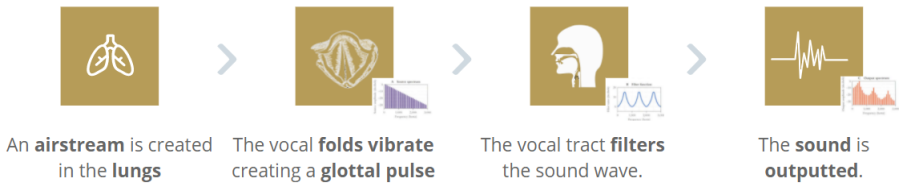

Figure 2.1: Sound production and filtering processes.

Components with the highest amplitude dominate the speech content - *formants* - and carry most of the spoken content of the signal. The lowest frequency of these components is the fundamental frequency and directly impacts the speaker's perceived pitch.

Sambur [14] identifies fundamental frequency (pitch or F0) and formant frequencies (F1, F2, F3, and F4) as two of the most informative acoustic features for speaker identification. Indeed, formant frequencies and energy are instrumental in modeling specific speech traits in the pronunciation of nasal phones, vowels, and strident consonants. Sambur [14] states that individuals have prominent patterns in the vocal tract filtering process, directly impacting the typical phones' pronunciation.

Conversely, fundamental frequency depends on the vocal cords' thickness and height; hence is particularly useful for gender and speaker identification [14]. Men tend to have an average fundamental frequency average around 125 Hz, whilst women average 210 Hz – i.e., a significant discrepancy that dominates the gender perception of vocal traits.

2.2 Voice Profiling

Speech interface technology, which includes automatic speech recognition, synthetic speech, and natural language processing, is beginning to have a significant impact on business and personal computer use. As a result, there is a growing need for machine learning (ML) systems that are capable of classifying, detecting and extracting information from speech signal.

In the particular case of speech applications, the process of deducting personal characteristics and information about the circumstances and environment of a speaker from their voice is called *profiling from voice* [13]. Typical profiling methods use signal analysis techniques by computing audio features – compact and accurate mathematical representations of the sound. Indeed, audio features can be thought not only as particular characteristic of the signal, but also, when aggregated, as a proxy for actual vocal traits of the speaker. Therefore, each audio feature has a specific usage and utility when it comes to capture specific characteristics of the speaker (such as, age, emotional state, gender, etc.).

For the purpose of our research, we are mostly interested in two research areas in the speech field that deal with speaker profiling: speaker identification, and gender categorization through voice. Accordingly, the following subsections outline the state-of-the-art audio features for each of these tasks.

2.2.1 Speaker Identification Through Voice

While several proposals of vocal traits taxonomies have been developed [13, 11], the work of Sharma et al. [15] provides an updated and application-oriented framework for acoustic features. The author proposes a division into six different categories: time-domain features, frequency-domain, time-frequency, wavelet, cepstral, and deep features. However, for the purpose of speaker profiling, Sharma et al. [15] highlight four of the previous categories: time-domain, frequency-domain, cepstral and deep features. Below are synthesized the most commonly used audio features for the purpose of speaker identification for each of the four selected categories.

2.2.1.1 Time-domain features

Time or temporal domain features measure properties of the signal throughout time. Given that all sounds correspond to a time series signal, time-domain features are the simplest way of analyzing a signal in its original form. Time-domain analysis is particularly straightforward for signals that are either short, or stationary over time. Speech signal is, however, non-stationary. To address this, most common temporal features cut the signal in short chunks of quasi-stationary signal, using windowing techniques, and analyze the feature distribution across consecutive chunks of the signal. For the purpose of speaker identification, there are three most frequent subgroups of temporal features: amplitude-based, energy-based and rhythm-based features.

Amplitude-based features investigate the signal's temporal envelope, namely its fluctuation along

time, being shimmer its most frequent measure. Shimmer computes cycle-to-cycle variations of the amplitude in a waveform. It is commonly used as an input for speaker recognition and speaker verification systems.

As the name states, energy-based features measure the level of energy carried by the signal and its variation through time. The signal's energy contour provides information on the signal's spoken content and the speaker — their affect, emotional and psychological state, health, and various other factors. Given that the energy carried by the signal is typically variable in time, energy-based features (such as RMS) capture the energy values over consecutive intervals of time. Loudness or volume is also a relevant energy-based feature, which is mathematically defined as the root mean squared value of the signal's magnitude within a frame.

Finally, rhythm-based features capture regular patterns in the speech signal. Most common rhythm-based features are speech duration, articulation rate, phoneme duration, pause ratio, total duration, total pause duration, total vowel duration, pulse metric, and speaking rate [14].

2.2.1.2 Frequency-domain features

The time-domain shows the signal variation throughout time. Yet, in such domain, the signal is analyzed in an aggregated fashion, hence with a very low granularity. To address this, using auto-regression or Fourier transform, the time-domain signal can be converted into a frequency-domain signal, allowing us to decompose a signal in its frequency components. Features created over in this domain called of *frequency-domain* or *spectral features*. Sharma et al. [15] group spectral features into five major sub-groups: STFT, chroma related features, auto-regression, tonality, and spectrum-based features. For the purpose of speaker profiling, Sharma et al. [15] highlight two major sub-groups of spectral features: tonality and spectrum-based features.

Tonality features capture the tone and intonation of the speaker in an utterance. The four most common tonality based features are Fundamental Frequency (F0), jitter, frequency formants (F1, F2, F3, F4), and *Harmonic-to-Noise Ratio* (HNR). F0, commonly known as pitch, is the lowest frequency of a periodic waveform and captures a tone's degree of highness or lowness. Following up on pitch, we find jitter, which captures pitch variations across consecutive speaking periods. Typical applications include speaker recognition, and age/gender estimation systems.

Frequency formants capture concentrations of acoustic energy around particular frequencies in the speech wave. They provide information on typical speaker's vocal tract patterns, which makes them reliable features for the purpose of speaker identification. Finally, HNR separates noise from the harmonic part of the signal, i.e., tones' sounds. The ratio between those two parts is the Harmonic-to-Noise Ratio (HNR), which is often used to estimate the level of hoarseness of a voice.

Spectrum-shape features characterize the distribution of energy across the frequencies of the signal. There are various types of spectrum-based features, namely: spectral centroid (*brightness* of the signal), spectral spread (deviations of energy around the centroid), spectral skewness (level of deviation from the normal distribution), spectral kurtosis (flatness of the spectrum), and spectral

slope (how quickly the spectrum tails off towards the high frequencies).

Finally, auto-regression-based features are extracted from linear prediction analysis of a signal. Linear Predictive Coding (LPC) [16] is one of the most important features based on auto-regression. LPC removes redundancy from a signal and attempts to determine the following values by linearly combining the previously known coefficients. LPC based features are used mainly for audio retrieval, and segmentation.

2.2.1.3 Cepstral features

Cepstral features, also known as cepstrum, are obtained by taking the inverse Fourier transform of the logarithm of the spectrum of the signal. There is a complex, power, phase, and real cepstrum. Power, however, is the most accepted cepstrum representation in speech signal processing [17], using measures such as MFCCs, LPCCs, and PLPs. The cepstrum features are used primarily in speech recognition, pitch detection, speaker recognition, and speech enhancement.

Mel-frequency cepstral coefficients - MFCCs are one of the most frequent acoustic representation for speech signal, having wide application in speech recognition systems. MFCCs provide a representation of the human hearing perception since they filter and transform the original signal to mimic the way humans perceive it. They are a good representation of the short-time power spectrum, which can ultimately can be used to represent the vocal tract shape. As stated in Section 2.1.1, the vocal tract shape includes the articulators such as teeth, nasal cavity, and tongue, which together filter the sound impulse, thus generating the speaker's voice's resonance. Therefore, this shape can give a precise illustration of the phoneme being formed if controlled precisely.

MFCCs are originated from audio cepstral representation and represent a sound frame by a vector with 39 elements, from which we typically use the 13 first coefficients. This shorter representation includes 12 cepstrum coefficients plus the energy term. The remaining features correspond to the delta and the double delta, which characterize feature changes over time and provide the context information of a phone. Algorithm 1 outlines the extraction process of MFCCs from signal.

Algorithm 1 MFCC extraction algorithm

1. **Result**: Mel frequency cepstrum coefficients
2. Input: Audio Signal X
3. Frame the signal into short frames use windowing.
4. For each frame, calculate the periodogram estimate of power spectrum.
5. Apply the mel-filter bank to power spectrum, sum the energy in the filter.
6. Take logarithm of filter-bank energies.
7. Take discrete cosine transformation (DCT) of the log filter-bank energies

Perceptual Linear Predictive (PLPs) [18] use the same *rationale* as MFCCs, yet introduce changes in the MFCCs pipeline, namely during the calculation of cepstral coefficients (PLPs use LPCs instead of linear prediction algorithms) and the signal filtering (by not considering Mel filter

bank). Literature also makes available alternative cepstrum representations, such as LPCCs (Linear Prediction Cepstral Coefficients) [19], RASTA-PLP (Relative Spectral Transform PLP) [20], and GTCCs (Gamma-Tone Cepstral Coefficients) [21]. These representations use a similar algorithm to the one above, yet introducing small changes on the considered auditory scale for the transformation.

2.2.1.4 Deep Features

Deep Neural Networks (DNNs) mark a new phase in automatic speaker recognition technology evolution. They provide a powerful way to extract highly discriminating speaker-specific features from recording the speech [22]. The obtained features – deep features – can be extracted from various levels of the network, thus representing different granularity levels.

State-of-the-art speaker recognition systems use deep features to create speaker embeddings, a fixed-size vector representing the vocal traits of the speaker. The two most common examples are x-vectors [22] and i-vectors [23], which use DNN and UBM encodings, respectively. X-vectors can be seen as the replacement of i-vectors, which were the previous standard for speaker recognition systems. I-vectors consist of a universal background model (UBM) and a large projection matrix T that are learned in an unsupervised way to maximize the data likelihood [22]. The projection maps high-dimensional statistics from the UBM into a low-dimensional representation, which output the so-called i-vector[23]. Recently, ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification) [24] embeddings are becoming the state of the art in speaker identification. They add the attention mechanism and are more robust to noise and channel conditions.

In spite of being highly-descriptive, deep features show a limited potential to be verified and interpreted as they are retrieved from black-box models. Accordingly, they are commonly used as a raw input for other AI models that study the individual's voice, namely gender and speaker identification algorithms [17].

2.2.2 Voice Gender Perception

Voice perception can be thought of as a mixture of low-level perceptual feature extraction and higher-level cognitive process. Ultimately, in the specific case of voice gender perception, such exercise can be reduced to a match between the perceived vocal traits and a predefined representation of what an individual from a specific gender group typically sounds like, i.e., a gender model.

The identification of the vocal traits that most contribute to the definition of vocal gender models is a subject that has been investigated for the past few decades. Indeed, Pernet and Belin [25] identified two high-influential vocal traits to the gender perception process: pitch and timbre. The authors proved that pitch is used only when timbre information is ambiguous (i.e., for more androgynous voices) and that the sole use of pitch for classification allows to obtain good results for classifying gender.

In general, pitch shows consistent differences across gender groups: male adults tend to have voices with lower pitch when compared to the female counterparts. Moreover, pitch has also been proved to be inversely proportional to the height of the individual, i.e., the taller the speaker, the lower his pitch tends to be. Timbre, on the other hand, reflects the mixture of harmonics and their relative height, enclosing all-vocal traits that cannot be qualified as pitch or loudness. Indeed, timbre is often described as the set of characteristics that gives color and personality to the voice, i.e., the unique attributes on the speaker's voice.

Combining these two vocal dimensions, Pernet and Belin [25] conclude that the ability to perceive gender can be mediated by vocal acoustical properties such as pitch, formant values (F1, F2, F3), glottal function, and spectral slope.

As a last note, little research has investigated the vocal patterns of individuals who do not identify as men or women and instead identify with non-binary genders. Despite such scarceness, recent research [26] identified that transgender men and women tend to pattern according to their gender identity, rather than biological sex, in terms of both vocal pitch and intonation characteristics. Hope and Bradley [26] suggest that non-binary individuals produce their pitch and vary their pitch in ways that are different from those with binary identities. Despite these findings, the authors alert for the danger of generalizing such results since the group of people identifying as non-binary is more heterogeneous and, consequently, shows a higher variability of vocal traits.

2.3 Speech recognition systems

As already mentioned in Chapter 1, one of the growing means of interaction with our systems is through voice. Voice-user interfaces (VUI), like *SIRI*[3] and *Alexa*[4], make spoken human interaction with computers possible, using speech recognition to understand spoken commands, and text-to-speech to reply. Indeed, several speech-based commercial applications exist nowadays, ranging from home automation systems, virtual smartphones' assistants, or voice interaction systems in automobiles. Regarding these systems, we are particularly interested in automatic speech recognition systems (ASR), which will be the focus of our research.

Accordingly, in this section we start by covering the major ingredients that compose the ASR pipeline. Then, we introduce the some of the most frequent evaluation measures for speech recognition systems, and, finally, review the state-of-the-art toolkits for implementing and training ASR systems.

2.3.1 ASR pipeline

The key focus of automated speech recognition systems (hereinafter ASR) is finding the most probable word sequence for an observed audio chunk. In other words, the system finds the word sequence W with the highest likelihood given the observed feature vectors X. Mathematically, we can model this either in the discriminative or the generative approaches:

$$\begin{aligned} &Word\ sequence : W = w_1, w_2, ..., w_m \\ &Acoustic\ observations : X = x_1, x_2, ..., x_n \\ &W^x = \underbrace{arg_w max[P(W|X)]}_{\text{discriminative model}} = \underbrace{arg_w max[P(X|W) * P(W)]}_{\text{generative model}} \end{aligned} \tag{2.1}$$

For the past few decades, speech recognition has mostly been on a generative approach. These models learn how to match audio signals to words, generating the instance space composed by all possible word sequences. Then, using *decoding* techniques, the algorithm searches for the most probable sequence of words.

The high-level architecture of generative speech recognizer can be deconstructed as a set of sequential building blocks. Indeed, they allow us to go from an observed audio signal to phones, then to sequences of phones (word), and finally to a sequence of words. Figure 2.2[5] maps the high-level framework of these systems.

The basic framework of speech recognizer systems comprises three major stages: capture, transducing, and decoding. The transducing phase corresponds to the ensemble of three different

[3]https://www.apple.com/siri/

[4]https://developer.amazon.com/en-US/alexa

[5]https://jonathan-hui.medium.com/speech-recognition-series-71fd6784551a

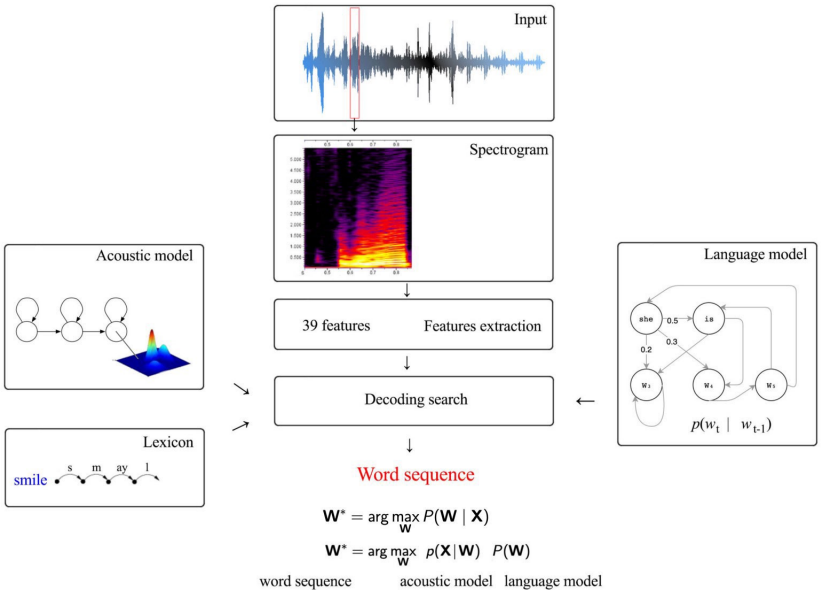

Figure 2.2: Basic framework of a generative speech recognizer system.

models: acoustic, lexicon, and language model. Joining these pieces, we obtain an end-to-end ASR framework which follows the steps below:

1. **Extract information on the observed signal** - starting from an audio clip, we use sliding windows to extract acoustic features. Each window will generate a sequence of vectors, one for each frame.
2. Transduce the **acoustic features** of the signal to obtain all possible words sequences. The transduction process corresponds to an ensemble of three different models:
   (a) **Acoustic Model** - $P(X|W)$ - used to recognize phones from the obtained acoustic features.
   (b) **Lexicon Model** - used to identify the most probable sequence of phones. In short, it will transform phones into words.
   (c) **Language Model** - $P(W)$ - used to identify the most probable sequence of words - utterances.
3. Combine the information provided by the three previous models using an **ASR decoder**, that will guide the search in the space of all possible word sequences.

We will briefly cover each of the above blocks in the following sections. It is worth noting that we will only cover essential aspects of an end-to-end generative speech recognizer. Most recent frameworks may include additional or distinct elements that will not be discussed in the following sections.

2.3.1.1 Acoustic Features

In speech recognition, knowing how we hear is more important than knowing how we speak[6]. Following this principle, the input of speech recognition systems aims to represent the speech signal's human perception. Indeed, as stated in Section 2.2, MFCCs are the accepted acoustic features to use as input for ASR systems.

To extract audio features, we conventionally use sliding windows of width 25ms and 10ms apart to parse the audio waveform. For each sliding window, we extract a frame of audio signals. We apply Fourier Transform, and manipulate it to make the perceived speech features stand out. Then we apply the inverse Fourier Transform. Several pre-processing techniques can be applied at this stage, being the most frequent ones PLP and MFCC. In the end, we extract a vector of features for each frame, which will be used as the input for the ASR system. As an example, X in $P(X|W)$ can be thought as the vector containing the MFCC values.

2.3.1.2 Speech Transducing

The generative approach looks for all possible sequences of words (with limited maximum length) and finds the one that best matches the input acoustic features. Therefore, this process allows us to go from acoustic features to words.

Acoustic models use MFCC features as an input to identify the most probable phones in the observed signal. Lexicon models use phones to find the most likely words (phone sequences). Finally, language models use words to find the most likely phrases (sequences of words).

2.3.1.2.1 Acoustic Model

Acoustic models estimate the likelihood of an audio feature vector X given a phone, i.e., $P(X|phone)$. They provide a powerful method to measure the distance between our observed audio frame and the typical MFCC representation of a phone. Phones, however, are dependent on the context, i.e., on the adjacent phones.

Articulation depends on immediately adjacent phones (*co-articulation*). Indeed, sounds change according to the surrounding context within a word or between words. With that in mind, when building a complex acoustic model, we should not treat phones independent of their context. The label of an audio frame should include the phone and its context–*triphones*. Indeed, as shown in Figure 2.3, the spectrogram for phoneme */eh/* varies with the context.

Two common modeling techniques for Acoustic models are Gaussian-Mixture models (GMM) and Deep Neural Network (DNN). GMM combines n Gaussian distributions, each with a specific weight, to form a new probability density function (an n-component GMM). Concerning DNN approaches, we can find three significant deep network possibilities for acoustic models: Fully connected (FC), Convolutional (CNN) and Recurrent (RNN) Neural Networks.

[6]https://jonathan-hui.medium.com/speech-recognition-series-71fd6784551a

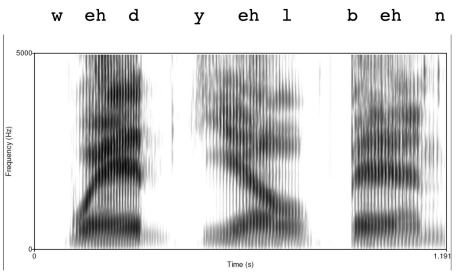

Figure 2.3: Three context-dependent spectrogram representations of the phone /eh/.

Fully connected (FC) networks directly use the Mel filter bank's features as an input to the deep network, which will encode a phone representation and its respective likelihood. Some FC models [27] contain 3–8 hidden layers with 2048 hidden units in each layer. Hence, this model can predict the context-dependent states' distribution from the audio frames. However, FC networks are computationally intense. It requires many model parameters, even for reasonable feature size. CNN takes advantage of locality and discovers local information hierarchically. On the other hand, time-delay neural networks (TDNN) explores the fact that audio speech is time-sequence data. Instead of applying a 2D convolution filter, we use a 1-D filter to extract features across multiple frames in time.

Finally, RNN is a deep network designed for time-sequence data using an LSTM mechanism. Long short-term memory (LSTM) is a type of recurrent neural network (RNN) where each cell has the input, the previous state, and the memory. Because of this architecture, LSTM is particularly useful to process sequences of data and it is used in some of the most accepted ASR frameworks like Mozilla's DeepSpeech [27].

2.3.1.2.2 Lexicon Model

Pronunciation lexicon models the sequence of phones of a word using Finite State Techniques (FST), frequently represented by Hidden Markov Models (HMM).

An HMM comprises hidden variables and observables and is modeled by the transition - likelihood of transiting from one internal state to another - and the emission probabilities - the likelihood of an observable given an internal state. Combining these two, we obtain the *forward probability*, which will express the likelihood of our observations. Then, we use decoding to find the most probable internal state sequences that matches our observations.

For the purpose of speech recognition, the observable is the content in each audio frame, and the internal state represents the phones identified in the observed acoustic features. The emission probability (representing the observable for each internal state) will be modeled by the acoustic model.

However, phones show changes in frequencies' amplitudes from the beginning until the end. To reflect that, we sub-divide the phone into three states: the beginning, the middle, and the ending part of a phone. To handle silence, noises, and filled pauses in a speech, we can label them as

SIL and treat it like another phone. We can also introduce *skip arcs*, arcs with empty input (), to model skipped sounds in the utterance, i.e., phones that are often mispronounced or omitted in dialogues.

2.3.1.2.3 Language Model

Even if the audio clip is not grammatically perfect (eg. contains skipped words), we assume that our audio clip is grammatically and semantically correct. Therefore, if we include a language model $P(W)$ in decoding, we can improve the speech recognition.

The language model deals with the likelihood of the word sequence. Since it estimates a sequence, it is no surprise that HMM is also used to represent language models. Hence, HMM language models work similarly to lexicon models: instead of predicting phones, they estimate the likelihood of a sequence of words. This creates an n-gram language model. Higher-order models ensure a greater granularity in our predictions.

The combination of the lexicon and acoustic models gives us $P(X|W)$ element for our generative speech recognizer – the likelihood of an observation in a space composed of all possible words. Introducing the language model - $P(W)$ - we obtain the likelihood of a sequence of words for a given recording - $P(W|X)$.

These three pieces, however, need to be combined in a efficient and organized way. To do so, ASR decoders are used to search for the optimal word sequence in a non-exhaustive way.

2.3.1.3 ASR Decoding

Speech recognition architectures commonly give the run-time decoder the task of combining and optimizing transducers (acoustic, lexicon, and language models). In this section, we will focus on Weighted Finite-State Transducers (WFST), the most common decoding technique.

Weighted Finite-State Transducers (WFST) is one of the most efficient transducers composition technique. Transducers encode a mapping between the input and the output label sequences. We start with an HMM transducer **H** to transform HMM states into context-dependent phones. By composing other transducers, namely the context-dependency (**C**), the pronunciation lexicon (**L**), and the grammar transducer (**G**), we map phones into a grammatically-sound sequence of words. The combination of these models leads to a decoding graph like the one in Figure 2.4.

The primary efficiency gain of WFST is in the composition of the H, C, L, and G transducers to form a single decoding/search graph, like the one presented below. However, we do not directly compose them together since the decoding graph would be too big to store or perform a full search. Transducer composition provides us with an optimization step, in which pruning is a vital part of the decoding. Most common prunning techniques are Beam Search, A* Search (Best-first search), and the Multipass search [28].

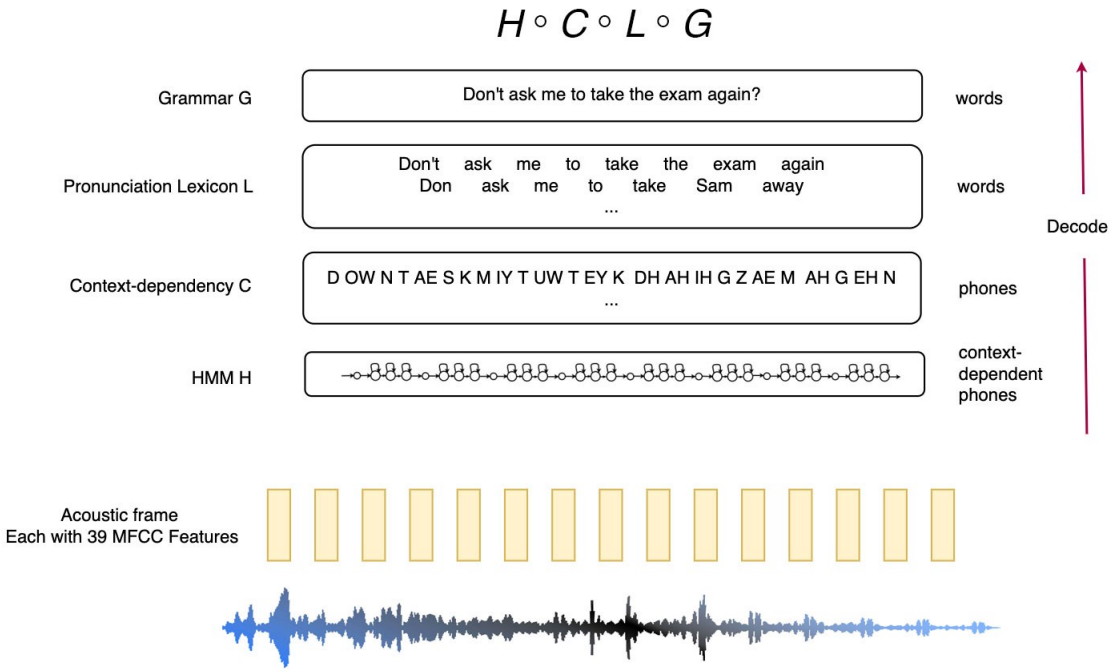

Figure 2.4: Basic WSFT pipeline, with the $H \circ C \circ L \circ G$ composition.

2.3.2 Evaluation

To evaluate speech recognition systems, the ASR system's output (*hypothesis text*), is compared to a literal transcription of input audio (*reference text*). Standard measures used in speech recognition evaluation are:

- **Word Error Rate** measures how many substitutions, insertions and eliminations - *edit distance* - are needed to convert the prediction to the true result - *ground truth*. The same formula can be applied on different levels: phoneme – Phoneme Error Rate (PER) –, character – Character Error Rate (CER), and Sentence Error Rate (SER).

$$WER = \frac{Insertions + Deletions + Substitutions}{Word\,Count} \tag{2.2}$$

- **Word Accuracy** measures the total number of correct words compared to the total number of words.

$$WAcc = 1 - WER = \frac{Number\,Correct\,Words}{Word\,Count} \tag{2.3}$$

2.3.3 State-of-the-art ASR toolkits

Until a few years ago, state-of-the-art for speech recognition was a phonetic-based approach including separate components for pronunciation, acoustic, and language models. Typically, this consists of n-gram language models combined with Hidden Markov models (HMM). Most accepted toolkits using this approach are HTK, and Kaldi.

Hidden Markov Model Tool Kit (HTK)[7] is used to build Hidden Markov Models (HMM) and can also be used in the designing of speech recognition system. HTK provides scripts for acoustic modeling, which can be changed for any other recognition applications. These tools uses HMM for training, testing and results analysis.

Kaldi[8] is a state-of-the-art automatic speech recognition (ASR) C++ toolkit, containing almost any algorithm currently used in the industry. It contains not only pre-trained models ran over popular datasets (such as Wall Street Journal Corpus [29] and TIMIT [6]), but also allows the user to train their own acoustic models. Accordingly, Kaldi provides tremendous flexibility and power in training own acoustic models and forced alignment system. The acoustic models are created by training the models on acoustic features from labeled data, or any other transcribed speech corpus.

Nonetheless, with the increasing usage of speech as an interface, major companies joined the industry by offering new architectures and tools for these systems. Prevalent examples come from NVIDIA (Neemo toolkit), Facebook (wav2vec toolkit), and Mozilla (DeepSpeech toolkit).

NVIDIA NeMo[9] is a Conversational AI toolkit powered by NVIDIA. The toolkit is an accelerator, which helps researchers and practitioners to experiments with complex neural network architectures. Speech processing (recognition and synthesis) and Natural Language Processing are the significant capabilities of the platform. The framework relays on PyTorch as the Deep Learning framework.

Facebook is another company with a strong presence in the speech industry. Indeed, Wav2vec was made available in 2019 as an extension to the open source modeling toolkit fairseq[10], and was announced as an important tool to provide better audio data representations for keyword spotting and acoustic event detection. Alongside wav2vec, Facebook showcased a new self-supervision model — ConvLM — that achieves state-of-the-art performance in correctly recognizing words outside of its training lexicon, and a lightweight sequence-to-sequence (seq2seq) model for speech recognition that's reportedly more efficient than previous work while delivering a better WER.

Finally, Mozilla launched the DeepSpeech [11] initiative in 2014, a simple, open, and ubiquitous speech recognition engine. Simple, in the sense that the engine should not require server-class hardware to execute. Open, in the sense that the code and models are released under the Mozilla Public License. Ubiquitous, in the sense that the engine should run on many platforms and have

[7] https://htk.eng.cam.ac.uk/
[8] https://github.com/kaldi-asr/kaldi
[9] https://developer.nvidia.com/nvidia-nemo
[10] https://github.com/pytorch/fairseq
[11] https://deepspeech.readthedocs.io/en/v0.9.3/index.html

bindings to many different languages.

The architecture of the engine was originally motivated by that presented in [27], and it is implemented over Google's TensorFlow toolkit. However, the engine currently differs in many aspects from the engine it was originally motivated by. The core of the engine is a recurrent neural network (RNN) trained to ingest speech spectrograms and generate English text transcriptions.

The DeepSpeech architecture is significantly simpler than traditional speech systems, which rely on laboriously engineered processing pipelines. This framework does not need a phoneme dictionary, nor even the concept of a *phoneme*. Key to DeepSpeech's approach is a well-optimized RNN training system that uses multiple GPUs, as well as a set of novel data synthesis techniques that allow us to efficiently obtain a large amount of varied data for training. Deep Speech also handles challenging noisy environments better than widely used, state-of-the-art commercial speech systems.

The standard architecture of the model uses 5 units (4 ReLU + 1 RNN). The first three are ReLU layers, and the fourth one is an RNN, which includes a set of hidden units with forward recurrence. The fifth (non-recurrent) layer takes the forward units as inputs. The system also uses a standard softmax output layer, and CTC (Connectivist Temporal Classification) beam search decoding.

2.4 Crowdsourcing data collection

As the industry expands to more natural forms of interaction with everyday devices and services, such as communication via natural language, the need for data to train such applications increases. Indeed, data-driven applications and intelligent systems, such as personal assistants or autonomous vehicles, require large amounts of data to train such systems. Furthermore, the wide range of new diverse and heterogeneous users demands for robust and unbiased solutions that perform successfully regardless of their individual characteristics or demographics.

To answer these needs, crowdsourcing emerged as an attractive solution for large data collection efforts, allowing to reach a diverse crowd in a expedite and scalable manner (both faster and less costly). The concept of crowdsourcing leverages the so-called *wisdom of crowds*, the idea that a large group of individuals can together provide surprising insight or value, even if individually they are inaccurate [30]. Therefore, crowdsourcing can be thought as a strategic model to attract a motivated crowd of individuals. These individuals, henceforth referred as contributors, can perform micro-tasks that take anywhere from a few seconds to several minutes to complete. Such tasks, Human-Intelligence Tasks (HITs), are typically made available via an online platform that distributes jobs across large numbers of people in exchange for a reward. Common tasks approached with crowdsourcing are labeling images, translating or transcribing text, or recording speech, to name a few.

Regarding the collection pipeline, Botelheiro et al. [8] suggests that the participation in these platforms, from the contributor's point a view, can be divided into four phases:

1. **Registration** - contributors sign up, providing demographic (age/gender) and language data (including reading, writing, and speaking proficiency per language).
2. **Work Selection** - based on their self-reported qualifications, the platform matches the contributors to a pool of tasks, which are organized as *Jobs*. A Job is a set of Human-Intelligence Tasks (HITs) with a common goal, such as "Record yourself reading the following sentences in English USA". Once the contributors accept the assigned Job, they are referred to as *Job Members*;
3. **Execution** - Job Members read the instructions of the Job and perform the HITs.
4. **Payment** - upon successful completion, payment is given to the contributors, proportional to their participation.

In the specific case of speech data collection, the collection pipeline entangles two steps: 1) a *generation step* in which the contributor is assigned with a prompt to read, and 2) a *validation step* in which the contributor is requested to validate certain aspects of previously recorded audio. Recordings that fail to meet the quality criteria are marked to return to the pool of available work in order to be re-recorded. Successfully validated recordings are then packaged for delivery, along with any additional relevant metadata (with respect to the recordings – such as the text that was recorded – or the speakers – such as gender or age).

The aforementioned validation step is an answer to new set of challenges of crowdsourcing that are caused by the loss of control by data requesters in a remote context, ultimately impacting the

quality of the collected data.

There are, however, two different motivations for the referred issues. On the one hand, the required input in a micro-task may be subjective or even ambiguous, leading to misalignment between the output and the instructions of the task. On the other hand, the reward behind each micro-task can cause contributors to minimize their effort, rush the work, or even attempting to cheat the system to get the reward without any effort. The latter is commonly known as *crowd frauds*. Typical *crowd frauds* involve a mismatch between the speaker and its self-collected stats (e.g., gender, age, etc.), potentially leading to quality issues in the obtained data. Most common crowd frauds are of three types:

1. **Gender mismatch**: the self-assigned gender of the speaker in the platform is incorrect.
2. **One account-many speakers**: a recording job contains multiple speakers,
3. **One speaker-many accounts**: the same speaker has multiple accounts in the platform.

Typical validation mechanisms are manual tasks where the Job Member is asked to validate certain aspects of the recordings (such as recording/noise conditions, nativeness and the match of the audio with the prompt). Given the sensitivity of this decision, each recording is validated by multiple contributors to ensure higher consistency and certainty in the final decision. Depending on the task, the volume of low-quality contributions may be considerable, and manually reviewing micro-tasks may take as much or more time and effort than performing them. Therefore, crowdsourcing platforms are trying to use AI in the process, namely by introducing speaker and gender identification algorithms in the collection pipeline.

However, different aspects of the data present different validation challenges. For instance, validating that a recording has no background noise is more straightforward than validating that the speaker is of a specific gender. Indeed, self-reported speaker information (provided when signing up for the platform, such as gender or age) is sensitive, and ultimately hard to verify and contest. Hence, self-reported stats on the speaker, like gender, are vulnerable to ill-intentioned contributors, ultimately impacting all tasks that rely on this data.

2.5 Bias and Fairness

Despite not having a universal definition, an unfair AI algorithm is frequently described in literature as one whose decisions are skewed toward a particular group of people [31]. Indeed, because machines can treat similarly-situated people and objects differently, research is starting to reveal some troubling examples in which the reality of algorithmic decision-making falls short of our expectations. As a result, some algorithms run the risk of replicating and even amplifying human biases, particularly those affecting misrepresented groups [2].

Prevailing examples come from facial recognition systems biased against black-skinned users[12], or even hiring algorithms that systematically downgraded the score given to women's *resumés*[13]. Naturally, as these conflicts are starting to be unveiled, companies and governmental institutions are starting to be held accountable for their models' behaviors and performance, especially with respect to minorities and marginalized groups.

As an answer to these events, new constraints and needs were introduced in AI. A major example comes from AI systems trained to help on high-stakes decisions in loan applications. To prevent the existence of bias conflicts in these sytems, the European Central Bank (ECB) introduced restrictions on the information that can be used to train these models, for instance by forbidding the usage of specific information the clients such as age and gender[14].

While highly related, the concept of *bias* differs from *fairness*. For that reason, in this section we start by distinguishing the two concepts. Then, focusing on speech applications, we will review the major bias conflicts unveiled in literature, while covering the most frequently employed mitigation techniques. Finally, we present some considerations on the effectiveness of such interventions, particularly in a crowdsourcing data collection scenario[15].

2.5.1 Bias $\neq$ Fairness

Broadly, fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making [1]. Fairness is thus a concept from Sociology that depends on individual perceptions, individual expectations, and context. Indeed, this subjectivity of the concept itself is one of the main reasons for it to be hard to achieve.

Verma and Rubin [32] identifies twenty different fairness definitions, while Mehrabi et al. [1] simplifies this approach and solely considers ten different definitions. Regarding the latter approach, we can group the prevalent fairness definitions into three major categories:

1. **Individual Fairness.** Give similar predictions to similar individuals [33].
2. **Group Fairness**. Treat different groups equally [33].

[12] https://www.wired.com/story/can-apples-iphone-x-beat-facial-recognitions-bias-problem/
[13] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
[14] https://www.ecb.europa.eu/paym/coll/risk/ecaf/html/index.en.html
[15] Details on crowdsourcing data collection available at Section 2.4.

3. **Subgroup Fairness**. Subgroup fairness is a mixture of the two previous fairness types. It picks a group fairness constraint like equalizing false positive and asks whether this constraint holds over a large collection of subgroups [34].

A fair system is one that verifies two major sets of conditions. First, a fair system should only act based on domain-specific (and ethical) information; and, secondly, its performance should be comparable across distinct classes of its user base. Certainly, the violation of one of the referred premises leads to unfair systems, which may be the result of poorly developed AI. Causes of such conflicts can then be drilled-down to each of the key components of AI systems: inappropriate methodology (algorithms, features, testing, monitoring) and biased data.

On the contrary, bias is a mathematical concept commonly defined as a systematic error against specific sub-groups. Indeed, it is a measure of favoring/hindering something and, if not controlled, may lead to unfair treatment of misrepresented groups. Bias in algorithms can emanate from unrepresentative or incomplete training data or the reliance on flawed information that reflects historical inequalities. If unchecked, biased algorithms can lead to decisions with a collective, disparate impact on specific groups of people even without the original intention to discriminate. Yet, if controlled, it can be a compensation mechanism hence fair.

Bias can be manifested in data in many shapes and forms. Indeed, Mehrabi et al. [1] identifies twenty different types of bias. For the purpose of our research, we will solely focus on two of the previous:

1. **Representation bias.** derives from inappropriate definition or sampling of the population [35], leading to diversity issues in the obtained training sets.
2. **Aggregation bias.** happens when false conclusions are drawn for a subgroup based on observing other different subgroups or when false assumptions about a population are taken [35].

Lee et al. [2] point out that bias can creep in during all phases of a project. This is usually the result of an unintentional emergent property of the algorithm's use rather than its programmers' conscious choice. Indeed, it can be challenging for developers to identify the problem's source or explain it to a court. While there are many causes, we focus on two: historical human biases and incomplete or unrepresentative data.

Pervasive prejudices shape historical human biases against certain groups, leading to their reproduction and amplification in computer models [2]. Suppose African-Americans are more likely to be arrested in the U.S. due to historical racism or other inequalities within the criminal justice system. In that case, these patterns will be mirrored in the training data and captured by the ML algorithm. If historical biases are factored into the model, it will make the same kinds of wrong judgments that people do [2].

Insufficient training data is another cause of algorithmic bias. Consider that an algorithm's training set is more representative of some people than others. In that case, the system will most likely show systematic errors against unrepresented or under-representative groups. Lee et al. [2] argues that it is often the lack of diversity in the training sample that leads to the under-representation of a particular group or specific attributes. Indeed, the latter, *data* or *input bias*, is the focus of

our research. To fight this, the author suggests that pre-processing tasks should be employed to correct the lack of diversity in the training set, ensuring a uniform distribution over one or more variables of the training set.

General methods for bias mitigation involve interventions before or during the model's training, and *fair interventions* over already trained models. Foster et al. [31] proposes a taxonomy that divides mitigation methods in three different levels: pre-processing, in-processing, and post-processing [1]. For the purpose our research, we will focus on the pre-processing level.

Pre-processing interventions transform train data to remove underlying discrimination. This approach intervenes at a design level by assuring an uniform distribution over one or more variables in the training set. Such variables can thus be thought as the balancement criterion of these interventions.

The premise explored by these interventions is that assuring a similar representation of groups in the training set guarantees a similar performance for the referred groups. The effectiveness of such interventions is, however, dependent on the ability of the selected variables to represent the complete spectrum of diversity in data. Whether it concerns speech, images, text, or other types of data, the more diverse a train set is, the more capable will the algorithm be to handle extreme and diverse scenarios. Therefore, the selection of the balancement criterion is an highly-impacting decision to take before balancing the training set.

Indeed, as the industry expands and demands robust and unbiased solutions, balancing the training set for AI systems has become a common practice among developers. Most common balancement criterion follow a *direct recipe*: once bias is detected against specific social sub-groups, then the training sets are adjusted to ensure an equal representation of over the same sub-groups in which bias is detected. As a result, descriptors such as gender and racial groups are some of the most frequent balancement criteria used in the AI industry.

Such variables are, however, general descriptions of individuals, i.e., mere proxies for actual characteristics on the individual. Indeed, if the referred proxies are misaligned with the actual profile of the individual, these proxies can be dangerous in the sense that they may be perpetuating offensive social stereotypes (for instance, what a male voice is expected to sound like, or what is the skin tone for an individual of a given racial group).

To fight this issue, there has been a growing effort to use verifiable and actual descriptions of individuals in the dataset. A prevalent example comes from ImageNet[16], an image database commonly used to generate train sets for facial recognition systems. As reported in the State of AI Report 2020[17], the company identified offensive categories, such as racial and gender expression characterizations, among ImageNet's database. Depictions such as race were complemented by descriptions of the skin tone of the individual, allowing developers to produce algorithms that more fairly classify faces and activities in images.

[16] https://www.image-net.org/
[17] https://www.stateof.ai/

2.5.2 Speech bias

Research has been conducted concerning possible bias conflicts in speech applications in recent years. As a result, speech bias has been identified against social groups of our society, such as gender [5], age [9], race [10] groups.

Garnerin et al. [5] identified bias against women in the performance of speech recognition systems by analyzing the gender representation in four major corpora of French broadcast. The authors concluded that the disparity of available data for both genders caused performance to decrease on women. Vipperla [9] identified that the ASR word error rates (WER) for older adults are significantly higher than those of younger adults. Bias was also identified against racial groups. According to Koenecke et al. [10], black speakers have an average WER ten points greater than white speakers when using state-of-the-art speech recognition systems — developed by Amazon, Apple, Google, and Microsoft [4]. Additionally, most of the example above were revealed to be caused by caused by input bias [10], i.e. the collection of speech data was not diverse enough, causing performance differences against the misrepresented groups.

As an answer to this, data providers' most prevalent interventions intervene on a pre-processing level, by assuring diversity in their datasets, i.e., variability. To do so, datasets are adjusted to ensure an uniform distributions over the same aspects in which bias is detected, most frequently across gender groups. Notably, an AI system's quality depends significantly on the volume and quality of the data used in its training. Data providers believe that by assuring a similar representation of genders in the training sets, they will observe an equal performance between gender groups.

The effectiveness of this technique is then dependent on two major premises: 1) the availability of accurate speaker metadata, and 2) the criterion applied for balancing data data must be diverse enough to capture the complete diversity spectrum of speaker profiles.

Concerning the first premise, as further detailed in Section 2.4, particularly in situations where the collection is not done in person (such as in a crowdsourcing platform), the obtained data is often vulnerable to ill-intentioned contributors [8]. Ultimately, such frauds lead to quality issues on the collected data, particularly affecting the self-reported speaker metadata. So, the metadata used on bias mitigation interventions shows the potential to have quality issues, hence to be inaccurate.

Focusing now on the second premise, data providers commonly balance datasets by gender labels. However, the employed criterion is not descriptive of the speaker, but a mere proxy of the vocal traits he is trying to emulate. In the sense that such representations generalize what an element of a given social group is supposed to sound like, they are only efficient in representing extremely homogeneous. Gender groups, however, are not homogeneous. Nor all male voices are low pitched, nor all female voice are high pitched. As a result, gender proxies show the potential to result in gender stereotypes, leading to misleading representations of individuals that contradict the dominant vocal traits of its gender group.

To illustrate this situation, as represented in Figure 2.5, we can think of a hypothetical situation in which one adjusts the training set to ensure a 50-50 distribution for female and male speakers. On an ideal scenario, this intervention would guarantee an equal of vocal profiles in the training

set. Yet, the actual scenario is a lot different: despite existing a dominant vocal profile within each gender group, the vocal traits distribution is not homogeneous. Indeed, there is a broader diversity spectrum for which the referred social proxies are relatively short in representing. The trained model could thus be biased against misrepresented or unrepresented vocal profiles. Most extreme examples rest on male's with high pitch voices, or even women's with low pitch voices could be systematically misrepresented simply because they fail to follow their sub-groups' typical vocal representation.

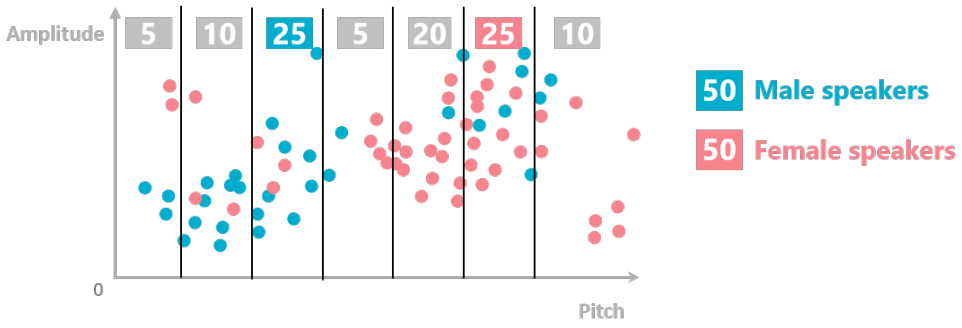

Figure 2.5: Proxy vocal traits representations: actual scenario.

The scenario presented above can enclose the types of bias referred in Section 2.5.1: aggregation and representation bias. Aggregation bias is the most evident example: false assumptions may be taken by generalizing the vocal profile for a gender group, hence neglecting the misrepresented individuals within the group. On the other hand, representation bias could be caused by balancing training sets across social groups. Given that these groups are way too broad, the prevailing voice representation would be based on the dominant group, potentially failing to represent minority groups and the complete diversity spectrum.

3 | Research Methodology

As algorithms drive more decision-making processes, machine learning models' tendency to learn our input data biases is a massive problem. Furthermore, the wide range of new diverse, and heterogeneous users demands robust and unbiased solutions that perform successfully regardless of their individual characteristics or demographics.

In the specific case of speech applications, research identified systematic errors against social groups of our society, such as female speakers, elderly speakers, or even misrepresented ethnic groups. To fight this, data providers' most prevalent interventions focus on assuring uniform distributions over the same aspects in which bias is detected, particularly across binary gender groups. However, balancing data along these features has three major drawbacks. First and foremost, as detailed in Section 2.4, these features are hard to test against when collecting audio for training such systems (particularly in a remote collection scenario). Secondly, they do not represent the individual's actual vocal traits (being only proxies of that). Finally, if used incorrectly, these proxies can be dangerous in the sense that they may be perpetuating social stereotypes (for instance, what a male voice is expected to sound like).

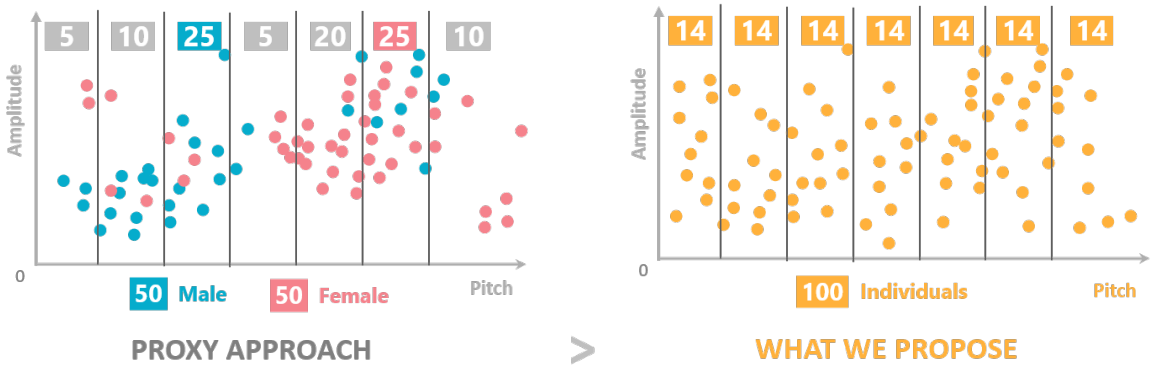

Figure 3.1: Research hypothesis: moving from proxy to actual vocal traits as balancement criterion.

To illustrate this, we can think of a hypothetical situation where a speech dataset containing 100 speakers is adjusted to ensure a 50-50 binary gender distribution. Figure 3.1 replicates this scenario by mapping the speaker distribution in the instance space using two vocal traits: pitch and amplitude. In an ideal scenario, this intervention would guarantee a similar representation of voice profiles in the dataset. The actual scenario is, however, a lot different: despite existing a dominant vocal profile within each gender group, the distribution of the vocal traits is not homogeneous. As one can see, there is a broader diversity spectrum for which gender proxies are relatively short in representing, which ultimately leads to a misrepresentation of speakers that

fail to follow the typical vocal profile of their gender group. In addition, as stated in Section 2.4, the considered gender stats are quite hard to verify and contest in a crowdsourcing context.

Given this background, this work explores the hypothesis of replacing proxies to the speaker's vocal traits (eg. gender), with actual vocal representations of the speaker to drive the data collection process. These traits should represent particularities of the speakers' voice (for instance, pitch) instead of proxies to the desired descriptors. It is important to highlight that measuring systems' performance across social groups (like the ones provided by gender information) is still relevant. However, this kind of sensitive self-reported metadata must not be contested based on normative (and potentially offensive) approaches, hence, they are not fit to drive data collection.

Our hypothesis would be represented by the second scatter plot in Figure 3.1, i.e., guiding the data collection using a representation that is blind to social groups and that uses an actual and verifiable criterion to balance data: pitch. This method not only covers a wider spectrum of diversity on the dataset (hence effective) but also ensures a similar representation of each profile in the data (hence fair).

Our research starts by identifying the vocal traits conveying the most information on the speaker, i.e., acoustic features that effectively differentiate individuals through voice. Then, we evaluate the impact of balancing such features in the training dataset for speech applications, analyzing the performance and bias impacts of these interventions. Our investigation was guided by the following two questions:

1. Which voice traits better differentiate and characterize speakers?
2. What is the impact (performance and bias) of balancing such features in training datasets for speech applications?

The following sections detail the defined methodology for each of the questions above, including the experimental setup and the considered data for each of the experiments.

3.1 Research Question 1

Research question 1 (hereinafter RQ1) explores the hypothesis of replacing proxy representations of the speaker with actual and measurable traits of his voice to drive the data collection process.

For this purpose, we divided RQ1 into two major phases. First, we start by identifying the vocal traits that best replace gender. Such analysis should offer a baseline set of vocal traits to replace the prevalent criterion for balancing speech data. Next, we go beyond gender labels and look for the vocal traits that best separate speakers in our dataset. Briefly, our research was guided by the following two questions:

1. Which observable vocal traits portray the same information as gender labels?
2. Which observable vocal traits best differentiate speakers?

3.1.1 Vocal Traits Conveying Speaker Characteristics

To identify the vocal traits conveying the most information on the speaker, the following four steps will be taken:

1. **Base Pool of Features** - define a pool of acoustic features from the Voice Profiling literature review.
2. **Utterance-level feature extraction** - extract the identified acoustic features for all recordings in our dataset.
3. **Speaker-level feature extraction** - aggregate the features' values at a speaker level, i.e., aggregate the values concerning all recordings of the same speaker.
4. **Selecting vocal features** - define a set of criteria to rank and exclude features from a pool of acoustic features.

3.1.1.1 Base pool of Features

The base pool of acoustic features was based on the Sharma et al. [15] taxonomy described in Section 2.2. In addition, the selected features will have to meet two different conditions: 1) verifiable and 2) semantically understandable.

Accordingly, this work will neither target deep features, nor cepstral features. Deep features are derived from black-box models, hence show a limited potential to be verified and interpreted. Regarding cepstral features, they are most commonly used as raw input for speech recognition systems and not for semantically understandable speaker profiling tasks. Briefly, cepstrum features are more helpful in representing an audio clip and its speech content than describing the speaker's traits in a concise and semantically understandable way.

As a result, we will work with two major groups of variables: time-domain and frequency-domain features. Concerning time-domain features, we will consider shimmer (amplitude fluctuations within the utterance), loudness (volume), energy (mean energy carried by the signal), and speaking rate (number of words per second). These four features should provide insights not only on the energy distribution in time but also capture the speaker's traits involving rhythm.

Regarding the frequency domain, we will use features from the two subgroups identified in Section 2.2: tonality and spectrum-shape. From the tonality group, we will use pitch (describing how high and how low a voice is), jitter (pitch fluctuations within the utterance), and HNR (Harmonic-to-Noise Ratio, a proxy for the level of hoarseness of a voice). Finally, we will analyze four spectrum-shape features: spectral centroid (a proxy for the brightness of a signal), spectral spread (average deviation around the centroid), spectral skewness (describing which regions of the spectrum concentrate the most energy), and spectral kurtosis (spectrum flatness around its mean value).

All things considered, we obtained an initial pool of eleven acoustic features: **spectral centroid**, **spectral spread**, **spectral kurtosis**, **spectral skewness**, **HNR**, **pitch**, **jitter**, **shimmer**, **loudness**, **energy** and **speaking rate**.

3.1.1.2 Utterance-level feature extraction

All features were extracted using the *Surfboard* toolkit [36], a Python package for audio feature extraction. Surfboard either calculates the value of the selected feature or, for variables that use sliding windows, computes a unique statistic on the feature. The toolkit thus provides for both cases a single feature value for each recording in our dataset.

In this work, we extracted the average value for all features and, whenever possible, also retrieved the standard deviation. Average should capture the feature's overall distribution, while standard deviation should evaluate how stable the selected measure is within the recording. The latter is relevant to identify cases for which mean is not representative, i.e., cases where the windowed instances are too disperse around the mean value for it to be significant.

3.1.1.3 Speaker-level feature extraction

The experiments conducted in RQ1 used scripted speech collections from DefinedCrowd's[1] proprietary crowdsourcing platform Neevo[2]. Considering that these are recordings of a single sentence, it is expected that the extracted features (on a utterance-level) show high variability, which ultimately reduces the significance of the obtained features. To mitigate this, we will aggregate the obtained features on a speaker level, i.e., to combine the features' values for all recordings of the same speaker.

To this purpose, three aggregation methods will be tested: mean, median, and trimmed mean. We are looking for the method that ensures the maximum inter-speaker variability, i.e., features that show the greatest differences between speakers. For this purpose, the coefficient of variation[3] is going to be used. The coefficient of variation (hereinafter CV) is a standardized measure of the dispersion of a probability distribution or frequency distribution.

Accordingly, to identify the aggregation method that maximizes the inter-speaker variability, we generated three different iterations of the train set, each using a different aggregation method. Then, for each train set and variable, we computed the total coefficient of variation and identified the aggregation method with the greatest variability for each variable.

[1] https://www.definedcrowd.com/

[2] https://www.neevo.ai/

[3] https://en.wikipedia.org/wiki/Coefficient_of_variation

3.1.1.4 Selecting the most promising candidates

Having identified and extracted our initial set of features (both at a speaker and utterance level), the final step in our pipeline is the selection of the most promising candidates. We are interested in features that are well-distributed across the instance space, i.e., with high variability. For this purpose, we once again used the coefficient of variation (CV)[1] to represent the variability of a given variable in a compact and unitless way.

CV, however, can be analyzed across multiple levels, namely *intra-utterance*, *intra-speaker*, *inter-utterance*, and *inter-speaker*. To reflect that, we defined a set of four premises to exclude the least promising audio features in each one of the considered levels. If any of the conditions are verified, the candidate will be excluded. The defined conditions are listed below:

- **Intra-Utterance**[4]: variables with very high variability within the utterance do not have a significant mean (aggregated) value and should be excluded.
- **Intra-Speaker**[5]: variables with very high variability across all recordings of the same speaker do not have a significant speaker mean value and should be excluded.
- **Inter-Utterance**[4]: variables with a very low variability across all recordings have low discriminatory power and should be excluded.
- **Inter-Speaker**[4]: variables with very low variability between speakers show a low discriminatory power and should be excluded.

There are no objective thresholds in literature to define a very high and very low CV value. Indeed, such thresholds are subjective and dependent on the variability of all variables in the dataset. Given this background, we set the following thresholds: $CV < 0.15$ identifies variables with very low variability, and $CV > 0.85$ identifies variables with very high variability.

3.1.2 Direct Replacement to Gender

The first research question investigates a direct replacement for gender models, i.e. vocal traits that can accurately emulate the speaker information carried by gender labels. To this purpose, our analysis was divided into two intermediate experiments: 1) measure traits correlation with gender, and 2) gender classification via traits.

The Spearman correlation[6] between gender labels and acoustic features captures the level of dependence between vocal traits and gender. Complementing these insights with non-parametric tests over the mean distribution of vocal traits between gender groups, we should obtain a shortlist of relevant features to replace the self-reported gender stats.

The next step in our analysis was to train a ML model to predict the speaker's gender from its vocal traits. The trained model follows an XGBoost architecture with the following parameters: *maximum depth = 3, random state = 1, number of threads = 4, evaluation metric = AUC, objective function =*

[4]Criterion only applicable to variables calculated using sliding windows.

[5]Criterion only applicable if the previous exclusion condition is not verified.

[6]https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

binary:logistic, early stopping rounds=10, and learning rate=1. The selected architecture was simplified to the maximum since our purpose was to assess the importance of each feature in an equal context, and not to obtain a top-of-the-line model for gender recognition. Considering this objective, the selection of the XGBoost architecture allowed us to not only obtain a significant performance (minimum of 80% accuracy), but also to have a high intelligibility on the obtained results. Indeed, using the feature importance tool from *xgboost* python toolkit[7], we should be able to identify the most informative vocal attributes for the identifying the speaker´s gender. Each feature will be scored by considering the mean information gain in the trained model.

3.1.3 Beyond Gender labels

Having identified direct replacements for gender labels, we moved our focus towards our second question: finding vocal traits not necessarily related to gender but still conveying valuable information to differentiate speakers. To do so, we divided our experiment pipeline into two steps. First, we identify the major correlation patterns in our pool of acoustic features. Then, using clustering algorithms, we determine which vocal traits better differentiate vocal profiles.

Clustering should allow us to obtain homogeneous groups of vocal profiles, and *post-hoc* analysis over the mean differences between clusters should capture the vocal traits that better explain between groups, i.e., variables with the greatest discriminatory power. To generate and analyze the obtained groups, we defined the following *clustering pipeline*:

1. Standardize the vocal traits' measures for all speakers.
2. Run hierarchical clustering and identify the optimal number of clusters (k). Cut the dataset into the optimal number of groups.
3. Calculate the mean vocal traits for the cluster.
4. Calculate the mean differences between each pair of clusters and identify the variables with the greatest standardized differences.

The pipeline above should allow us to identify the vocal traits carrying the most relevant acoustic features to differentiate vocal profiles in our dataset. Finally, comparing the obtained clusters with the self-reported gender groups should get us some insights on the effectiveness of the gender criterion to balance the dataset.

By the end of these two experiments, we should have a shortlist of vocal traits that can be used as criteria for balancing speech data. Accordingly, these insights will be used in our second research question, where we will evaluate the impact of balancing such features in the training dataset for speech applications, analyzing the performance and bias impacts of these interventions.

[7] https://xgboost.readthedocs.io/en/latest/python/python_intro.html

3.1.4 Dataset

The experiments conducted in RQ1 will use real data from DefinedCrowd's[8] proprietary crowdsourcing platform Neevo[9]. Our dataset consists of 155.9 hours of English as spoken in the United States of America, from a total of 275 native speakers. It comprises 63,363 recordings paired with the corresponding transcript and metadata on the speaker and the prompt's recording conditions. Each speaker has, on average, 2041 recordings in the dataset, totaling 34 minutes of recording time. The mean duration of the recordings in our dataset is thus 8.7 seconds. The distribution of the number Finally, to keep the focus on a solution that is solely dependent on the speaker, all entries in our dataset were recorded under a quiet environment.

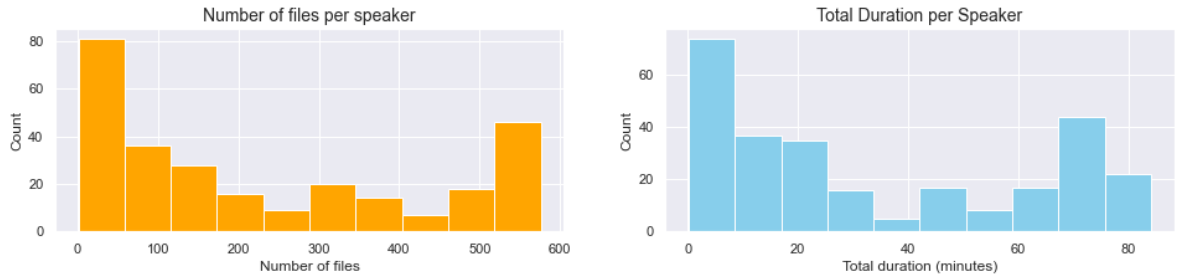

Figure 3.2: Distribution of speech time in the train set.

The speaker distribution across gender and age groups is not uniform in our dataset: it shows a prevalence of female speakers (64% vs. 36% of male speakers) and a significant concentration of speakers in the 20-40 age range. Such difference is reflected in the number of recordings and in the total recording time of each group. The most extreme example is the 50-60 age group, which only includes female speakers. The accents, however, show a greater diversity (42 different accents), with a subtle prevalence of the Californian accent. Table 3.1 summarizes the speakers' metadata profile.

All transcriptions were normalized using the jiwer[10] python package by applying a set of NLP transformations: lower case, contractions expansion, punctuation removal, removal of consecutive white spaces, and empty word tokens. Given that our analysis focuses on the speaker's vocal traits, we only considered recordings with a quiet background environment. All sentences containing characters not included in the English (USA) alphabet were excluded from our dataset.

The metadata on each file can be divided into two major groups: recording and speaker-related information. Our focus will be on the speaker, namely on his age, accent, and gender. Accordingly, our initial dataset includes both metadata and the selected acoustic features for each recording in our dataset. As thoroughly explained in Section 3.1.1.3, we will work over two different levels of analysis (utterance and speaker). Speaker, however, will be the primary level of analysis, which will imply an aggregation of the values of the features of all recordings of the same

[8]https://www.definedcrowd.com/
[9]https://www.neevo.ai/
[10]https://pypi.org/project/jiwer/

Characteristics	Speakers	
	Number	%
Gender		
Female	177	64.36
Male	98	35.64
Age		
(0, 20]	34	12.36
(20, 30]	106	38.55
(30, 40]	74	26.91
(40, 50]	37	13.45
(50, 60]	15	5.45
(60, 100]	8	2.91
Living Country		
Canada	1	0.36%
United States	268	97.45%
Not Provided	6	2.18%
Total	275	100.00

Table 3.1: Speakers' metadata profile.

speaker. All features were extracted using the *Surfboard* toolkit [36], a Python package for audio feature extraction. Surfboard either calculates the value of the selected feature or, for variables that use sliding windows, computes a unique statistic on the feature. The toolkit thus provides for both cases a single feature value for each recording in our dataset.

3.2 Research Question 2

Research Question 2 explores the hypothesis of using actual vocal representations of the speaker (such as pitch, and loudness) to ensure measurable balancing on speech data collections. For this purpose, we estimate the impact of balancing the training set of speech applications over a given setup of vocal features (hereinafter balancement criterion). The considered features will be the ones identified in RQ1 as the most informative to differentiate speakers through voice.

The area chosen for research is, however, extremely extensive, given not only the plurality of speech applications and the millions of utterances that state-of-the-art speech applications require for training. Therefore, we will narrow down our study to the impact on automated speech recognition systems (ASR).

To this purpose, using a common framework, several ASR systems will be trained, each with a specific distribution of vocal traits in their train sets. Indeed, using a given setup of vocal attributes (eg. pitch), we will identify reference groups in the data set and assure a distribution of speech data as uniform as possible for those groups. *Post-hoc* analysis on the systems' performance

should give us insights on the impact of the considered vocal traits in the performance of the ASR systems. To ensure comparability, all ASRs were trained under the same conditions (architecture, parameters, and number of hours in the train set). For this purpose, five steps were taken:

1. **Pool of Recordings** - define a pool of recordings from which we will extract the train sets.
2. **Feature extraction** - extract the predefined set of acoustic features for all recordings in the base pool.
3. **Train set generation** - create the train set, extracting a sample from the final pool of recordings with an uniform distribution across one of the selected balancement criterion.
4. **Model training** - train the model using a standard architecture.
5. **Evaluation** - evaluate and compare the performance of the models using a common test set and evaluation metrics.

The following sections detail the considered methodology for each of the steps above, and the considered data for all analysis in this chapter.

3.2.1 Pool of recordings

The pool of recordings serves as a repository of speech data to generate the different train sets needed for our experiments. Considering that each train set follows a specific distribution of vocal traits, the defined pool must be diverse and large enough to meet several criteria for balancing. Therefore, we set 500 hours as the minimum size for our final pool of recordings.

The experiments conducted in RQ2 will use the Common Voice Corpus 6.1 English [11] dataset (hereinafter CV dataset). CommonVoice [37] is part of Mozilla's initiative to help teach machines how real people speak, and it currently is one of the largest publicly available voice datasets of its kind. The voice clips are readings from a bank of donated sentences corresponding to the dictation/monologue product. Once the recordings are validated, they enter the dataset.

The original CV dataset comprises 1,686 of validated hours speech, from a total of 66,173 speakers. Each entry in the dataset consists of a unique MP3 paired with its transcription. Many of the recorded hours in the dataset also include demographic metadata like age, gender, and accent. Therefore, to obtain a more manageable pool of files, we filtered this dataset, using the pipeline presented in the figure below.

Further, since we analyzed the acoustic features on a speaker-aggregated level, we set a minimum and a maximum speaker time: 20 seconds and 30 minutes, respectively. Additionally, to overcome quality issues in the metadata files, we filtered the *CV dataset* by only considering transcriptions with English characters (a-z, '','' and '''''), and complete gender metadata. All transcriptions were normalized, using a predefined set of NLP transformations: lower case, contractions expansion, punctuation removal, removal of consecutive white spaces, and empty word tokens.

By the end of this process, we obtained an *Initial Pool* of files, from which we extracted two random samples of 30 hours, which served as *dev* and *test* set for all models trained. To ensure

[11] https://commonvoice.mozilla.org/

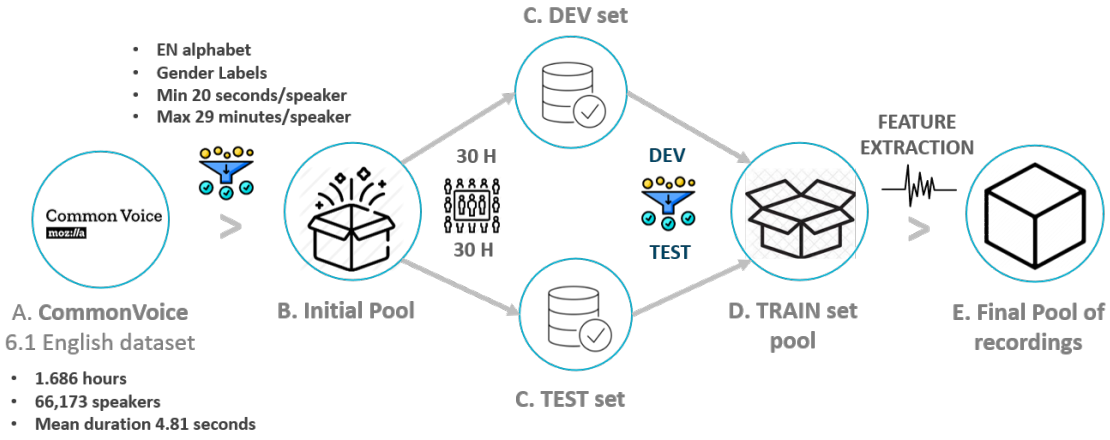

Figure 3.3: Filtering pipeline for our base pool of recordings.

that none of the recordings in these two sets were included in the train set, we excluded these 60 hours of speech data from our pool of recordings and obtained the *Train Set Pool* – including all recordings that can be used to form our train sets. This pool contains a total of 585.97 hours of English, from 13,471 native speakers. Each speaker has, on average, 30.7 recordings in the dataset, totaling 2 minutes and 28 seconds of recording time. The mean duration of the recordings in our dataset is thus 4.81 seconds.

The speaker distribution across gender and age groups is not uniform in our dataset: it shows a prevalence of male speakers (79.3% vs. 20.7% of female speakers) and a significant concentration of speakers in the 20-40 age range (63.6%). The majority of the individuals in our pool lives in the United States of America (36.2%), being then followed by England-based speakers. Table 3.2 summarizes the speakers' metadata profile.

This pool was then complemented with information on a predefined set of acoustic features, for each of the MP3 files contained in the dataset, as thoroughly explained in the following section of this document. By the end of this process, as represented in Figure 3.3, we obtained our *Final Pool of Recordings* which will then be used to extract the several train sets to use in our experiments[12].

[12]Further details on each of the obtained trains sets are made available in Section 5.1

Characteristics	Speakers	
	Number	%
Gender		
Female	2787	20.7
Male	10684	79.3
Age		
(0, 20]	1582	11.7
(20, 30]	5698	42.3
(30, 40]	2872	21.3
(40, 50]	1497	11.1
(50, 60]	952	7.1
(60, 100]	783	5.8
Blanks	87	0.6
Living Country		
Australia	465	3.5
Canada	593	4.4
England	1509	11.2
Indian	1226	9.1
USA	4878	36.2
Other	1021	7.6
Blanks	3779	28.1
Total	13471	100.00

Table 3.2: Speakers metadata for the Final Pool of Recordings.

3.2.2 Feature Extraction

The considered acoustic features will be the ones identified in RQ1 as relevant for identifying speakers through voice. Accordingly, all features will meet two different conditions: 1) verifiable and 2) semantically understandable. These features will then be used to as balancement criterion for the training sets of the ASR systems. They will, however, be divided in two major groups of variables:

1. **Direct replacements to gender** - variables identified as conveying similar information to gender labels. Systems that use these variables as balancement criterion will be compared with an ASR model that use *gender* as balancement criterion.
2. **Gender blind representation** - variables identified as being relevant for differentiating individuals through voice, but not necessarily related with the speaker gender.

The selected features were extracted on a utterance-level for all files contained in the *Train set pool*, and then aggregated at a speaker level. To this purpose, we considered the aggregation methods identified in Section 3.1. These methods were chosen with the purpose of maximizing the inter-speaker variability, i.e., to obtain features with the most differences between speakers.

All features were extracted using the same tools and methods identified in RQ1. Accordingly, our feature extraction tool is *Surfboard* toolkit [36], a Python package for audio feature extraction. Surfboard either calculates the value of the selected feature or, for variables that use sliding windows, computes a unique statistic on the feature – in our case, mean.

3.2.3 Train set generation

A crucial step in our experiment pipeline is the selection and generation of the train sets for the ASR systems. The considered train sets should have an uniform representation for a given setup variables (hereinafter, *balancement criterion*), i.e, a similar speech time for k fixed-sized groups.

The number of groups for balancing (hereinafter, *reference groups*) depends on the variables selected as *balancement criterion*. On the one hand, for categorical variables, the number of groups equals the number of distinct values of the variable. On the other hand, for continuous variables, fixed-sized discretization was performed. Therefore, the selection of the most promising train sets encloses two steps. We start by identifying vocal traits carrying relevant information on the speaker, hence the most promising *balancement criterion*. Then, for continuous variables, we study the optimal number of bins for each setup of vocal traits.

For the purpose of our research, we are interested in train sets with a total size of 200 hours (hereinafter *objective*), and a number of speech hours as similar as possible for all reference groups. Therefore, the maximum number of hours to include in each bin is given by the ratio between the *objective*, and the number of groups (k). This ratio (hereinafter *bin objective*) will be an important variable of our analysis since it sets our objective of hours for each bin in the train set.

We start our study, by setting a minimum of two and a maximum of ten bins for train sets balanced over a single vocal trait. For train sets balanced over two vocal traits (eg. pitch and jitter), our problem is a two-dimensional one. Considering k_1 and k_2 as the number of bins for the discretization of the two vocal traits used as balancement criteria, the total number of groups in the train set (k) will be given by $k_1 * k_2$ – the product of the number of groups for each of the features contained in the pair.

Fixed-sized discretization process was performed for all extracted features. Further, all features were transformed to a speaker-level, using the aggregation techniques identified in Section 3.1.1.3. The bins' ranges were calculated over a subset of the dataset, removed of any extreme outliers, i.e., outside the $1.5 * IQR$ threshold. Finally, our discretization algorithm also includes a relaxation mechanism that replaces the limits of the most extreme bins with 0, $-\infty$, or $+\infty$.

Having generated all possible train sets, we focused on ranking the obtained train sets by evaluating their **uniformity** –i.e., if all bins in the train set have a similar amount of data (measured by the number of hours of speech recordings). For this purpose, as represented in the equation below, we created the *UniScore*, given by the ratio between the total number of hours in the train set[13] and *objective* of hours in the train set. The score ranges between 0 and 1 – 1 correspond-

[13]The duration of each recording is measured in seconds. To convert these values to the hour scale, we multiplied the duration's values by a factor of 3600.

ing to a perfectly uniform train set. This measure should capture massive discrepancies in the distribution of hours in the train set, hence detecting the least promising candidates.

$$\text{UniScore } = 3600 * \frac{\sum \text{Duration}}{\text{Objective}} = 3600 * \frac{\sum \text{Duration}}{200} \tag{3.1}$$

This shortlist, however, may still include train sets with major discrepancies in the distribution of data across reference groups. Naturally, having such a misrepresented bin compromises our objective of obtaining a uniform distribution of data across several groups. To detect such cases, we defined a second filtering criterion: *bin uniformity* – how similar is the volume of data across reference groups. To this purpose, we created the *BinUniScore*, represented in the equation below, where k_{min} is the bin with the smallest size in the train set, and k is the number of reference groups considered. This measure is given by the ratio between the minimum number of hours per bin in the train set, and the *bin objective*. The score once again ranges between 0 and 1 - 1 corresponding to training sets where every bin is represented to its maximum.

$$\text{BinUniScore } = k * \frac{\sum \text{duration}_{\text{kmin}}}{\sum \text{Duration} * 3600} = k * \frac{\sum \text{duration}_{\text{kmin}}}{\text{Bin Objective}} \tag{3.2}$$

Accordingly, the *BinUniScore* should identify train sets containing misrepresented groups. Ultimately, if we consider that these misrepresentations stop us from reaching the hour objective for the train set, this score can be thought of as a measure of how easy one can achieve the 200 hours objective using a specific setup of vocal traits. Therefore, we are looking for datasets with a score as close to 1 as possible. To this purpose, we once again set a threshold of 0.8, which identified train sets with a maximum difference of 20% in the size of each bin in the dataset.

Accordingly, the *BinUniScore* should identify train sets containing misrepresented bins. Ultimately, if we consider that these misrepresentations stop us from reaching the hour objective for the train set, this score can be thought of as a measure of how easy one can achieve the 200 hours objective using a specific setup of vocal traits. Therefore, we are looking for datasets with a score as close to 1 as possible. To this purpose, we once again set a threshold of 0.8, which identified train sets with a maximum difference of 20% in the size of each bin in the dataset.

Finally, we will consider one last measure in our decision process: *MaxUniSize* – the maximum size of our train set, with a perfectly uniform bin representations. The equation below details the formula for such measure, where k_{min} is the bin with the smallest size in the train set, and k is the number of groups considering for balancing data.

$$\text{MaxUniformSize } = k * \sum \text{duration}_{\text{kmin}} \tag{3.3}$$

Once we apply the referred thresholds we should have a shortlist of datasets with a total size around our 200 hours objective and a maximum difference of 20% in the distribution of hours across bins. Therefore, this shortlist of train sets will be referred as the **200 hours group**.

Conversely, when relaxing the Bin and UniScore thresholds to 0.5, we should be able to identify train sets that meet the uniformity premises, but for a 100 hours objectives. Naturally, all elements included in the 200 hours groups are also included in this group. Therefore, we will name this group of datasets as the **100 hours group**.

Finally, the previous models will be compared with a gender-balanced and an unbalanced model (hereinafter, *non-vocal models*). Considering that we have two groups of train sets with large differences in their size (100 and 200 hours), we will only compare models in the same group. Therefore, we generated two versions for the train set of the non-vocal models: a first one with 100 hours, and a second with 200 hours.

3.2.4 Model training

The next step in our analysis is to train multiple ASR systems, each using a train set with an uniform distribution for a given setup of vocal traits. To ensure comparability, all ASRs were trained under the same conditions (architecture, training time, and the number of hours in the train set).

For this purpose, all systems will follow Mozilla's DeepSpeech architecture, implemented over Google's TensorFlow. The core of the engine is a recurrent neural network (RNN) trained to ingest MFCC's and generate English text transcriptions. This framework does not need a phoneme dictionary, nor even the concept of a *phoneme*. Key to DeepSpeech's approach is a well-optimized RNN training system that uses multiple GPUs, as well as a set of novel data synthesis techniques that allow us to efficiently obtain a large amount of varied data for training.

The standard architecture of the model follows the standard DeepSpeech recipe [27]: 5 hidden units (4 ReLU + 1 RNN). As represented in Figure[14] 3.4, the first three are ReLU layers, and the fourth one is an RNN, which includes a set of hidden units with forward recurrence. Finally, the fifth (non-recurrent) layer takes the forward units as inputs. The system also uses a standard softmax output layer, and CTC (Connectivist Temporal Classification) beam search decoding. Considering that the focus of our research rests on the vocal traits of the speaker, we did not train any language model and used DeepSpeech's pre-trained language model[15] for all models. Finally, DeepSpeech uses the Adam method [38] for training.

All systems were trained under the DeepSpeech architecture, using the following parameters: number of hidden layers (1024), learning rate (0.0005), and dropout rate (0.3)[16]. In addition to this, we used a test batch size of 64, a dev batch size of 16 and automatic mixed precision. Each model was trained over 100 epochs, with a early stopping mechanism with a minimum delta of 0.2, over 10 epochs.

[14]Figure taken from: https://deepspeech.readthedocs.io/en/v0.9.3/DeepSpeech.html

[15]https://deepspeech.readthedocs.io/en/v0.9.3/USING.html

[16]DeepSpeech's base values for each of the parameters are the following: number of hidden layers (2048), learning rate (0.001), and dropout rate (0.05).

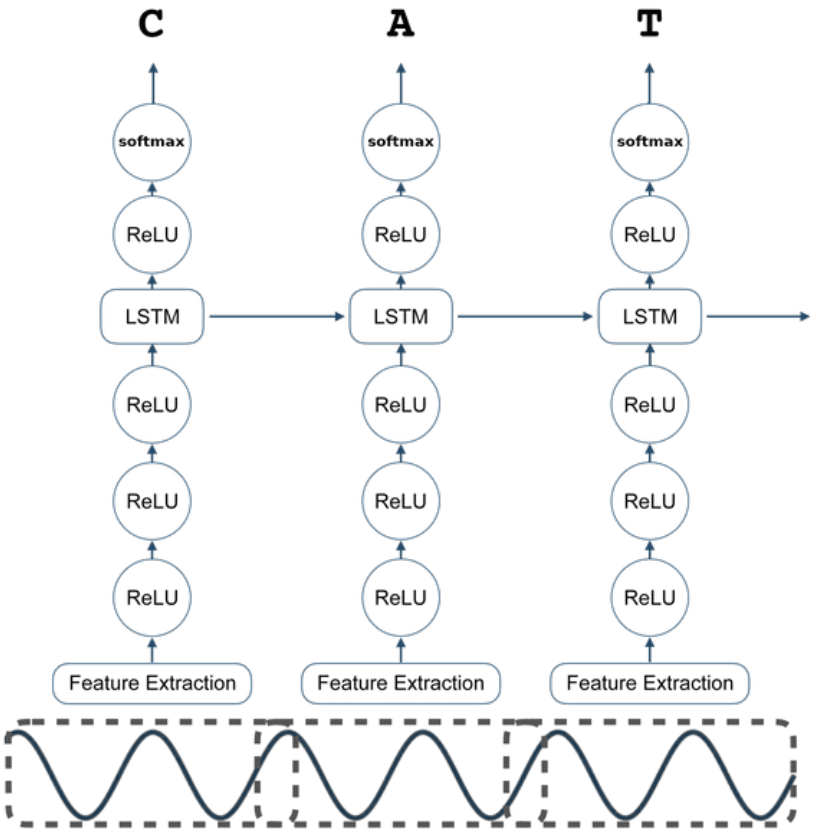

Figure 3.4: DeepSpeech's base architecture.

3.2.5 Evaluation

Concerning evaluation, Word Error Rate (WER) will be our primary performance metric. It is a measure indicating errors in alignment of text representation (actual vs. perfect) of audio, taking into account words omitted, inserted, or wrongly replaced.

The performance of each model was measured over three different test sets: 1) a 30 hours random sample of the CommonVoice dataset, 2) a 9.5 hours structured sample of the CommonVoice dataset where each speaker is represented by 5 different files, and 3) the LibriSpeech (test-clean) test set[17] – a 5.5 hours dataset used in literature as a benchmark for speech recognition architectures. Considering the two CommonVoice test sets, it is worth noting that they are independent from our training data, i.e., all speakers in the test sets were excluded from our *Train Set pool*.

At this point it is worth reminding the original purpose of this experiment: to assess the impact of vocal traits in the performance of speech applications. To this end, we will train train several ASR systems, each with a specific distribution of vocal traits in their train sets, and compare the performance and bias of the obtained models. Considering that our train sets will either have

[17] https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean

100 or 200 hours of data, which, according to Chuangsuwanich [39] – represented in Figure 3.5 –, should guarantee a word error rate (WER) around 45% and 40%, respectively. Given this 5 percentage points difference, we will only compare the performance of ASR systems trained with a similar amount of data.

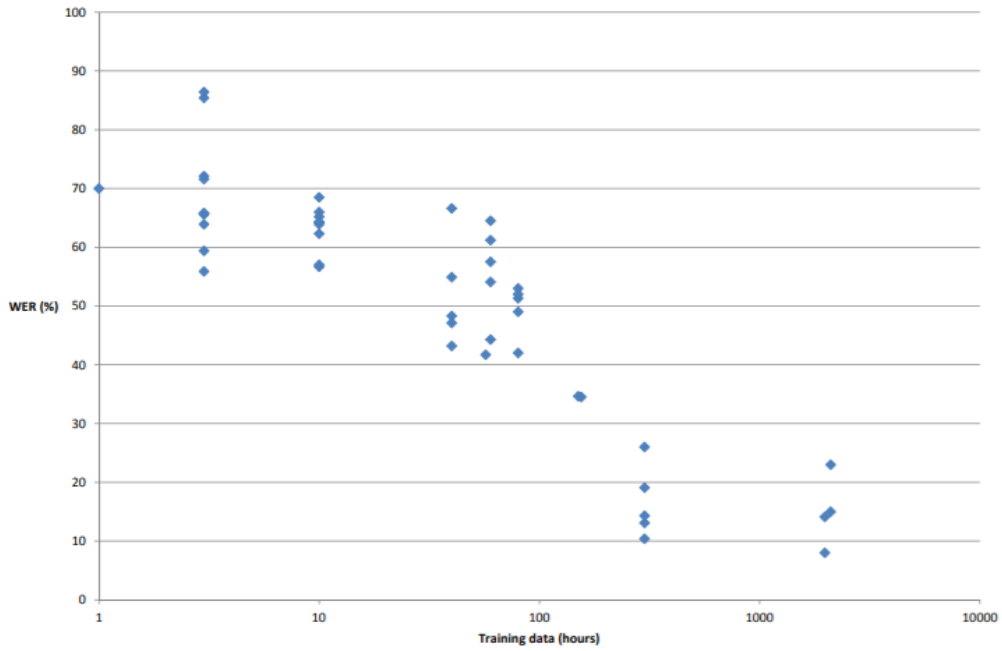

Figure 3.5: Expected WER with respect to amount of training data.

Instead of obtaining top-of-the-line ASR systems, we will focus on obtaining comparable systems, i.e., trained in a similar context (system architecture, train set size, and training time). Using a similar architecture, the only difference between models will be the distribution of vocal traits in the train set. Hence we should be able to evaluate the individual impact of a specific setup of vocal traits in the system´s performance and bias.

Therefore, our analysis will focus on detecting significant performance differences between ASRs, both on the global systems' performance, but also by comparing how biased each model is.

Performance will be compared using Wilcoxon pairwise tests [40] (α= 0.05), which will assess the existence of significant differences in the WER distribution in each of our test sets. To this purpose, we will consider all three test sets: the two CommonVoice test sets, and LibriSpeech. Despite independent from the training data, the two CommonVoice test sets are expected to have a better performance than the LibriSpeech set – the recording conditions in which the system was trained are similar to the two first test sets, hence impacting the effectiveness of the acoustic model. Finally, considering that we have three test sets, the weighted average of the WER in each test set (pondered by the number of hours words in each test set) was was our main performance measure.

Bias, on the other hand, will be evaluated by comparing the group performances over two metrics: age (teens, twenties, thirties, forties, fifties, Over60) and gender (male and female) groups. To this purpose, we will evaluate the existence of significant differences between groups, using either the Wilcoxon pairwise test (α= 0.05 – for gender groups, hence with $k = 2$), or the Kruskal-Wallis test (α= 0.05 – for age groups, hence with $k > 2$). The choice of these tests is motivated not only by the non-normality of our data (non-parametric tests), and by the number and (in)dependence of the groups for which performance will be compared. Given that LibriSpeech does not make available the required speaker metadata, our analysis will be limited to the two CommonVoice test sets (30 hours, and 9.5 hours with a balanced distribution).

By the end of this analysis, we should have insights on the impact of each vocal trait on the systems performance and bias.

4 | RQ1: From proxy to actual representations of the speaker

This chapter covers our first research question: *Which voice traits better differentiate and characterize speakers?*. We explore the hypothesis of replacing gender proxies with actual vocal representations of the speaker to drive the data collection process. To this purpose, we divided our RQ1 in two major phases. First, we start by identifying the vocal traits that best replace gender. Next, we go beyond gender labels and look for the vocal traits that best separate speakers in our dataset. Briefly, our research was guided by the following two intermediate questions:

1. Which observable vocal traits portray the same information as gender labels?
2. Which observable vocal traits best differentiate speakers?

The following sections of this chapter will the pipeline in Section 3.1. Accordingly, we start by defining a pool vocal traits the most information on the speaker. Then, Section 4.2 identifies the vocal traits that best emulate the information provided by gender labels. Section 4.3 finds the acoustic features that best differentiate speakers, while ignoring their dependency with gender. Finally, Section 4.4 discusses the obtained results and suggests the final set of vocal features to be used in RQ2.

4.1 Vocal Traits Conveying Speaker Characteristics

To identify the vocal traits conveying the most information on the speaker, as stated in Section 3.1, four steps will be taken. We start by identifying a base set of features from the literature (see Section 2.2). Then, we extract and aggregate the acoustic features on a speaker level, and finally, we select and rank the most promising acoustic features using a predefined set of premises.

The base pool of acoustic features considered in your analysis was based on the Sharma et al. [15] taxonomy described in Section 2.2. In addition, the selected features will must meet two different conditions: 1) verifiable and 2) semantically understandable.

Once applied these conditions, we obtained a base pool of eleven acoustic features: spectral centroid, spectral spread, spectral kurtosis, spectral skewness, HNR, pitch, jitter, shimmer, loudness, energy and speaking rate.

Having identified and extracted our initial set of features (on a utterance level), the next step in our pipeline was the aggregation of the obtained features on a speaker level. To this purpose, three aggregation methods were tested: mean, median, and trimmed mean. We are looking for the method that ensures the maximum inter-speaker variability, i.e., features that show the greatest

differences between speakers. For this purpose, the coefficient of variation[1] is going to be used.

Median proved to be the method with the best behavior for all variables, except for jitter, for which trimmed mean proved to be the best aggregation technique. Finally, given that the Surfboard toolkit extracted multiple Shimmer and Jitter implementations, we selected the implementation of each feature with the highest variability[2].

Finally, we moved to the selection of the most promising candidates. We are interested in features that are well-distributed across the instance space, i.e., with high variability. For this purpose, the coefficient of variation (CV) was used to represent the variability of a given variable in a compact and unitless way.

CV, however, can be analyzed across multiple levels, namely *intra-utterance*[3], *intra-speaker*[4], *inter-utterance*[4], and *inter-speaker*[4]. To reflect that, as further detailed in Section 3.1.1.3, we defined a set of four premises to exclude the least promising audio features in each one of the considered levels. If any of the conditions are verified, the candidate will be excluded.

Table 4.1: CV values for the base pool of acoustic features

Feature	Intra-Utterance	Intra-Speaker	Inter-Utterance	Inter-Speaker
Spectral Centroid	0.139	0.265	0.230	0.230
Spectral Spread	0.068	0.160	0.147	0.147
HNR	[5]	0.250	0.184	0.184
Pitch	0.104	0.234	0.220	0.220
Jitter	[5]	0.385	0.274	0.274
Shimmer	[5]	0.557	0.304	0.304
Loudness	[5]	0.256	0.234	0.234
Energy	[5]	0.859	0.772	0.772
Speaking Rate	[5]	0.558	0.200	0.200
Skewness	**1.388**	[6]	[6]	[6]
Kurtosis	0.045	**0.047**	[6]	[6]

Once applied the defined premises, we obtained the results presented in Table 4.1, from which we excluded two features from our initial pool: spectral kurtosis and spectral skewness. Spectral kurtosis showed a 1.388 intra-utterance score (hence an excessively high variability within the same utterance), while spectral skewness showed variability across recordings, hence reflecting a low discriminatory power.

All things considered, the initial pool of features was reduced to the following nine features to be analyzed on a speaker-level: **spectral centroid**, **spectral spread**, **HNR**, **pitch**, **jitter**, **shimmer**, **loudness**, **energy** and **speaking rate**.

[1] https://en.wikipedia.org/wiki/Coefficient_of_variation

[2] Detailed results for each variable and aggregation method available at Figure 7.1, in the Appendix Section.

[3] Criterion only applicable to variables calculated using sliding windows.

[4] Criterion only applicable if the previous exclusion condition is not verified.

4.2 Replacing Gender with Vocal Traits

The first research question of the present work investigates a direct replacement for gender models, i.e. vocal traits that can accurately emulate the speaker information carried by gender labels. To this purpose, our analysis was divided in two intermediate experiments: 1) measure traits correlation with gender, and 2) gender classification via traits.

4.2.1 Measuring traits correlation with gender

The Spearman correlation[5] between gender labels and acoustic features captures the level of dependence between vocal traits and gender. Complementing these insights with non-parametric tests over the mean distribution of vocal traits between gender groups, we should obtain a shortlist of relevant features to replace the self-reported gender stats.

Only three variables from our pool show strong correlation patterns with gender: pitch (-0.77), jitter (0.58), and HNR (-0.52). Such finding is reinforced when running the Kruskal-Wallis test[6] ($\alpha = 0.05$), which did not reject the hypothesis of existing mean group differences between gender groups for each of the three variables. Conversely, no other features in our pool showed dependency patterns with gender (both correlation and mean group differences).

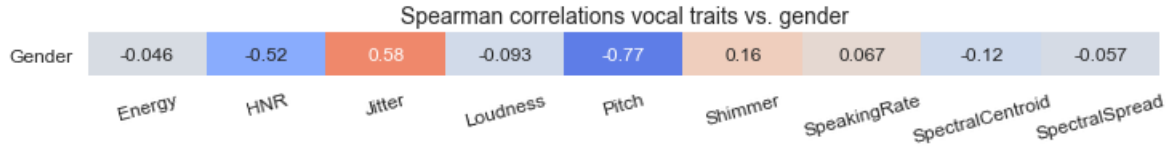

Figure 4.1: Spearman correlation between vocal traits and gender

Accordingly, the three variables above define our shortlist of relevant features, where we can identify two different levels of influence. As expected, pitch is the variable that most contributes to differentiate male and female speakers in the dataset. Not only it shows the highest correlation with gender, but it also has the greatest mean differences across gender groups[7]. On a second level of relevance, we find HNR and jitter, with medium-high correlation with gender (-0.52 and 0.58, respectively). Finally, it is worth noting that jitter and HNR are the two variables that most correlate with pitch (-0.65 and 0.59, respectively). Such finding is symptomatic of the vital contribution of pitch to identify gender models, which potentially spread its influence over its most correlated vocal traits, thus overrating the HNR and jitter contribution to explain gender vocal models.

Gathering the previous insights, we conclude that pitch is the most promising vocal trait to replace gender models, being then followed by jitter and HNR.

[5]https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

[6]https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php

[7]Box-plots containing the gender mean differences for each acoustic feature in the RQ1 shortlist can be found in Figure 7.3, in the Appendix Section.

4.2.2 Gender classification via traits

The next step in our analysis is to train an XGBoost model to predict the speaker's gender from its vocal traits. *Post-hoc* analysis on the contribution of each feature for the model should give us the most informative vocal attributes for identifying the speaker's gender. Each feature will be scored by considering the mean information gain in the trained model.

This pipeline was replicated over six iterations of our dataset, each containing a different combination of acoustic features. All six models were trained using a similar architecture: *maximum depth = 3, random state = 1, number of threads = 4, evaluation metric = AUC, objective function = binary:logistic, early stopping rounds=10, and learning rate=1.* The selected architecture was simplified to the maximum, since our purpose was to assess the importance of each feature in an equal context, and not to obtain a top-of-the-line model for gender recognition.

The latter architecture was implemented over six different iteration of our data set, each corresponding to a specific subset of our shortlist of acoustic features: 1) the complete dataset, 2) excluding pitch , 3) excluding pitch, jitter and HNR, 4) only pitch, 5) only pitch and jitter, and 6) only pitch and HNR. The accuracy results for each of the iterations can be found in Table 4.2.

Table 4.2: Gender Prediction - Test Accuracy

#	Iteration	% Test
1	Complete dataset	96.3
2	Excluding Pitch	81.7
3	Excluding Pitch + Jitter + HNR	70,1
4	Pitch only	87.3
5	Pitch + Jitter	89.1
6	Pitch + HNR	81.8

The first model (trained over the complete dataset) obtained a 96.3% test accuracy by considering all nine features in our pool. Regarding this model, when looking into the contribution of each feature, pitch revealed as the most informative variable (5.36 score). Following pitch, showing up once again in a second level of relevance, we find jitter with a 0.63 score. It is worth noting that the HNR score is quite similar to the one obtained by jitter (0.5 vs. 0.63), which once again sets these two variables on similar level of relevance for predicting gender through voice.

Having identified such a supremacy from pitch, we replicated the experiment pipeline over a train set deprived of information on pitch[8]. In such scenario, the test accuracy went down about 14 p.p (81.7%). Also in this model, we found jitter and HNR as the two most informative variables, with 1.78 and 1.40 scores, respectively. Therefore, in the absence of pitch, the model looks for indirect representations of the speaker's pitch, i.e., prioritizing information from its two most

[8]Feature contribution results for the models 2 – excluding pitch – and 3 – excluding pitch, jitter and HNR – can be found in the Appendix Section, Figures 7.5 and 7.4, respectively.

correlated variables. Finally, when training the model without any of the three variables in our shortlist (pitch, jitter, and HNR), we obtained a test accuracy of 70%, being shimmer the most influential variable for such prediction with a 0.77 score.

The obtained results for the three initial models confirmed the previously obtained insights: pitch is the most important vocal trait for the construction of vocal gender models, being then followed by HNR and jitter, both in a second level of relevance. Accordingly, to better understand their individual contribution for the prediction, we produced three additional models, each trained with a maximum of two acoustic features from our shortlist. Given the undisputed influence of pitch for the prediction, all combinations will include it. Accordingly, our final three models were trained over the following subsets of features: pitch, pitch + jitter, and pitch + HNR.

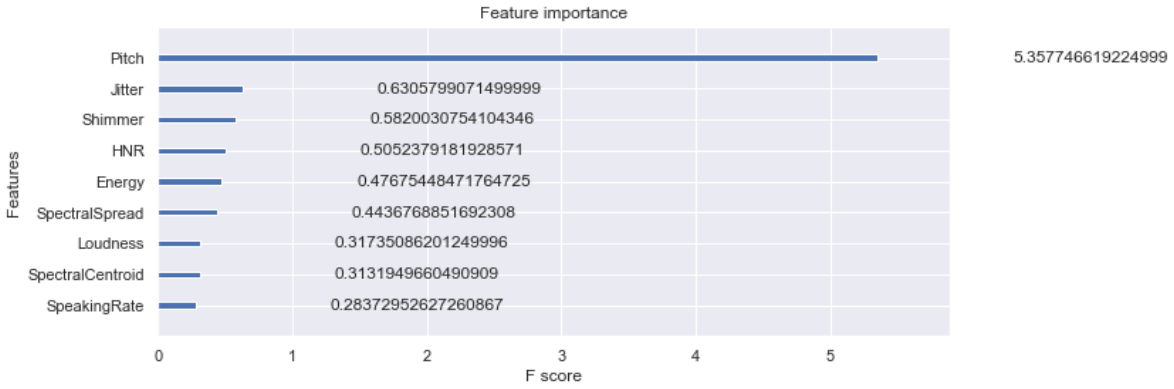

Figure 4.2: Feature contribution for the complete model to predict the gender of the speaker.

Pitch alone is capable of accurately predicting the speaker's gender for 87.3% of the individuals in our test set. Indeed, pitch alone is more effective in predicting gender than all the remaining eight variables combined (87.3% vs. 81.7%). To better understand the classification errors for the solution using only the pitch trait, its performance was measured through a grid search on the thresholds of 95% accuracy for the label "male" and "female" (using 5Hz steps). Through this process, the region between 128-148Hz was identified as having the most misalignment (27 speakers, 10% of the speakers). Indeed, such region contains 12 female and 15 male speakers, i.e., a 44%-56% gender distribution. For pitch values over 148 Hz, 94% of the speakers were females, and for pitch values under 128 Hz, 97% of the speakers were identified as male.

Such results[9] show us that pitch is quite efficient in separating gender, except for a grey area located in the 128-148Hz range, for which the gender distribution was close to perfectly balanced (143-148 Hz and 128-133 Hz showed a 50-50 gender distribution). Hence, 128-148Hz defines a *grey area* for which pitch is less competent in separating genders. This insight is consistent with the literature on voice gender perception, which state that the adult woman's average range is from 165 to 255 Hz, while a man's is 85 to 155 Hz [13]. The grey area seems thus to be located in the frontier of the two intervals.

At this point, it is worth reminding one of the original purposes of this experiment: obtain a small

[9]Detailed results can be found in Figure A, in the Appendix Section

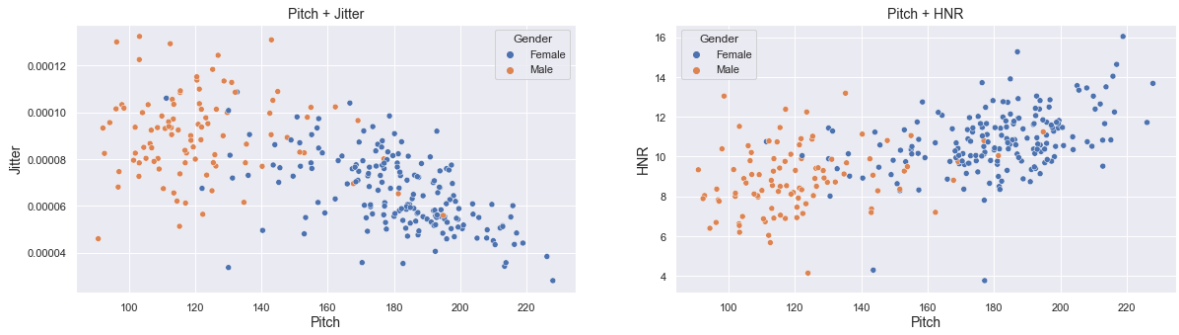

Figure 4.3: Scatter plots for the two pairs of gender relevant variables: Pitch + Jitter, and Pitch + HNR.

set of vocal traits that is accurate enough to replace the gender criterion when balancing training sets for speech applications. With that in mind, if we set an overly ambitious criterion, we may ultimately force data providers to exclude a large amount of valuable data, thus increasing the cost of the data collection jobs. Therefore, given this constraint, we defined a maximum of two vocal traits to be used as criteria. Our problem is now reduced to selecting the most informative pair of vocal features for gender recognition. Given the undisputed discriminatory power of pitch, all considered pairs of variables will include it. Again, this reduces our problem to the selection of one variable from the following group: HNR and jitter. The graphical representation of such hypothesis is presented in Figure 4.3.

Both scatter plots showed a good distribution across the instance space. Indeed, in both graphs, one can identify the 128-148 Hz pitch grey area, for which the gender separation across genders is not obvious. From a graphical point of view, jitter seems to be more effective at separating those individuals, by mapping male speakers to higher jitter values. Conversely, HNR appears to have a lower entropy for the speakers in that region. Thus, HNR seems to be less relevant for categorizing gender through voice.

These insights are also reinforced if we compare the test performance of the pitch solution (*pitch model*), with the two models that individually add information on jitter and HNR. When adding each of these two features to the pitch model, we obtained opposite results: jitter improved performance by 1.9 p.p (89.1%); whilst HNR decreased performance by 5.6 p.p. Accordingly, HNR seems to introduce noise into the prediction, generating a worse separation of the speakers in the instance space.

Gathering these insights, we are now in a good position to state that **pitch** and **jitter** are the **most promising vocal traits for replacing gender vocal proxies**.

4.3 Beyond Gender Labels

Having identified direct replacements for gender labels, we moved our focus towards our second research question: finding vocal traits not necessarily related with gender but still conveying valuable information to differentiate speakers.

To do so, we divided our experiment pipeline into two steps. First, we identify the major correlation patterns in our pool of acoustic features. Then, using clustering algorithms, we determine which vocal traits better differentiate vocal profiles.

4.3.1 Measuring the traits overall correlation

Correlation captures the degree to which two variables move in relation to each other. In this section, we start by looking into the Pearson correlation[10] between vocal traits and then identify the vocal features most correlated with gender (using Spearman correlation). Figure 4.4 presents the heat-map with the correlation values for all pairs of variables in our pool (namely the ones including gender).

Only two pairs of acoustic features show strong correlation: the *loud pair* (containing loudness and energy, with a 0.83 correlation) and the *spectral pair* (containing spectral centroid and spectral spread, with a 0.84 correlation). Such findings are, however, expected. Volume represents how humans perceive the energy carried by the sound wave. Hence it can be thought of as a transformation of the energy variable. Similarly, spectral spread captures deviations of energy from the spectral centroid.

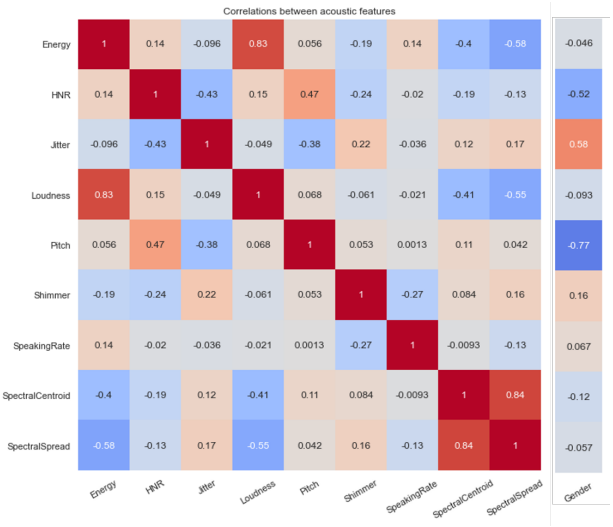

Figure 4.4: Correlation values for Gender-traits pairs, and traits-traits pairs.

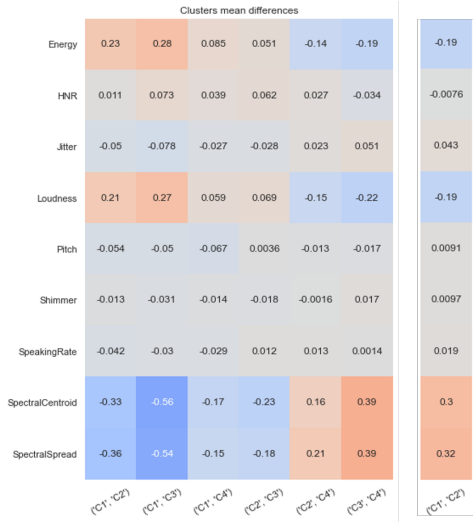

Figure 4.5: Standardized mean cluster differences for the complete dataset.

[10] ttps://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php

On a medium-high level of correlation, three major patterns were identified: 1) spectral centroid and energy (-0.4); 2) pitch, HNR, and jitter (each pair with absolute correlation values over 0.5); and 3) jitter and speaking rate (-0.44). It is worth noting that the second group of variables includes the three variables identified in Section 4.2 as the best replacements for gender vocal models. Additionally, these three variables also show the strongest correlation with gender: pitch (-0.77), jitter (0.58), and HNR (-0.52).

4.3.2 Differentiating vocal profiles

The next step in our analysis focus on identifying the vocal traits that best differentiate vocal profiles. To do so, we ran clustering over speakers in our dataset. *Post-hoc* analysis over the mean differences between clusters should capture the vocal traits that better explain between groups, i.e., variables with the greatest discriminatory power. To generate and analyze the obtained groups, we defined the following *clustering pipeline*:

1. Standardize the vocal traits' measures for all speakers.
2. Run hierarchical clustering and identify the optimal number of clusters (k). Cut the dataset in the optimal number of groups.
3. Calculate the mean vocal traits for the cluster.
4. Calculate the mean differences between each pair of clusters and identify the variables with the greatest standardized differences.

We started our analysis by looking into the obtained clusters for the complete dataset. The obtained dendogram[11] suggests two partition levels at $k = 2$ or $k = 4$, for which we computed the mean standardized differences between clusters. Figure 4.5 contains the obtained heat-map[12].

The two heatmaps show a common pattern: two pairs of highly correlated variables are responsible for the greatest mean differences between clusters. Spectral centroid and spectral spread show the greatest mean differences (0.3 mean absolute differences for $k = 2$), followed by energy and loudness (-0.19 mean absolute differences for $k = 2$). None of the remaining variables showed significant mean differences between clusters.

No gender pattern was identified for both partition levels (as an example, for $k = 2$, cluster one showed a 73F/34M gender distribution, while cluster two showed a 104F/64M distribution). This finding is consistent with the results presented in Section 4.2, since none of the three relevant variables for the gender voice perception (pitch, HNR, and jitter) show significant mean group differences between clusters (Kruskal-Wallis, $\alpha = 0.05$).

To validate these insights, we ran the *clustering pipeline* over four additional iterations of the original dataset: 1) a 50-50 gender-balanced sample, 2) the subset of male speakers, 3) the subset of female speakers, and 4) the subset of speakers located in the pitch grey area, identified in Section 4.2. For all four, the same two pairs of variables showed the greatest mean differences: the

[11] Dendogram available in Figure 7.6 in the Appendix Section.

[12] The obtained heatmaps for the subset of female speakers, and speaker in the 130-150 Hz pitch range can be found in the Appendix Section, Figures 7.8, and 7.7, respectively.

loud pair (containing loudness and energy), and the *spectral pair* (containing spectral centroid and spectral spread). Such results thus prove that these variables have a strong discriminatory power at different granularity levels and within each of the binary gender groups.

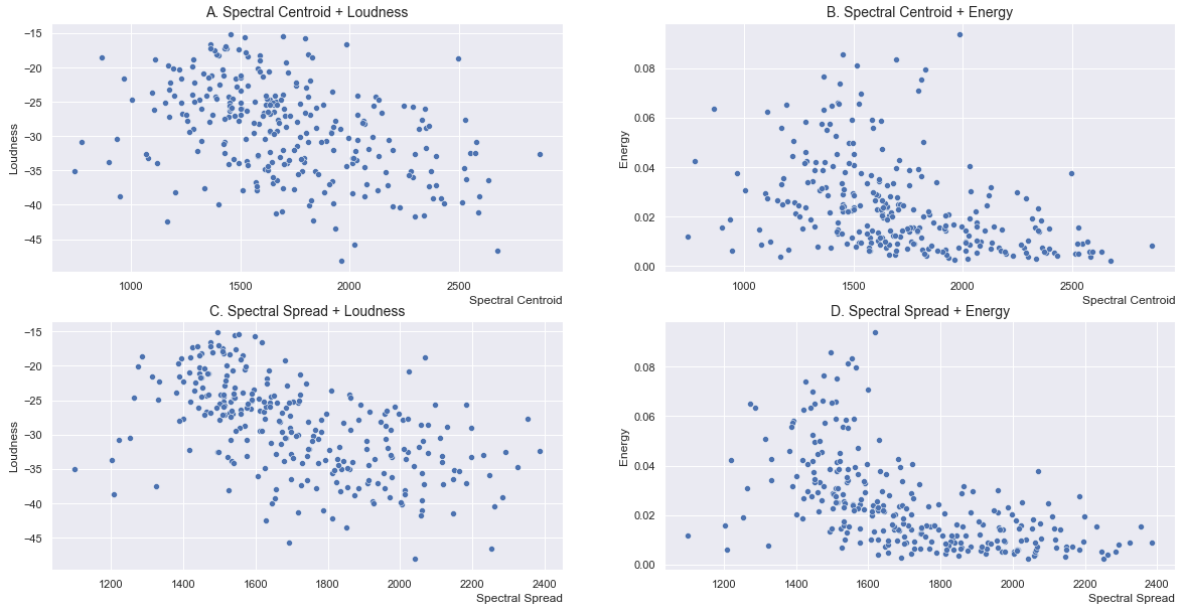

Figure 4.6: Four most informative pairs of variables for separating vocal profiles

Finally, we considered redundant to use more than one variable from each pair, given that both pairs contain highly correlated vocal traits. Therefore, our initial problem was now reduced to selecting the most informative pair of vocal traits from a shortlist of four hypotheses: A) spectral centroid and loudness, B) spectral centroid and energy, C) spectral spread and loudness, and D) spectral spread and energy. The scatter plots for each pair are available in Figure 4.6.

As one can see in the figure above, pairs A and C show the highest entropy, i.e., pairs of variables containing loudness seem better distributed across the space compared to pairs containing energy. Conversely, no graphical differences were found over the distribution of the two variables from the spectral pair: spectral centroid and spectral spread. To address this issue, we recovered the *inter-speaker* premise presented in Section 3.1.1.4: *variables with low variability between speakers show a low discriminatory power and should be excluded*. Once again, variability was evaluated using the coefficient of variation, being the obtained results presented in Table 4.1. Spectral spread showed the lowest inter-speaker variability in comparison with spectral centroid (0.147 vs. 0.23). Therefore, spectral centroid was the variable selected from the pair.

Gathering the previous insights, we are now in a good position to conclude that **spectral centroid**, and **loudness** are the two most informative variables to separate vocal profiles.

4.4 Most informative acoustic features

Our investigation was divided into two major phases. First, we identify the vocal traits that best replace gender. Next, we go beyond gender labels and look for the vocal traits that best differentiate speakers through voice.

Results show that pitch is the most relevant vocal trait to identify the gender of the speaker, being then followed by jitter. Indeed, pitch is quite efficient in predicting the gender in our data except for speakers located in the 128-148 Hz pitch range. For such interval, jitter provides important information to separate genders by mapping male speakers to higher jitter values.

Gender, however, may not be the most accurate representation of the speaker. Accordingly, our second research question looks for vocal traits not necessarily related with gender but still conveying valuable information to differentiate speakers. Our hypothesis is that the most accurate representation of the vocal traits of the speaker is not necessarily related with their gender. The obtained results indicate spectral centroid and loudness as the key variables discriminate vocal profiles - both weakly correlated with gender. Indeed, when running hierarchical clustering, the two explained most of the differences between the obtained groups. This finding was confirmed over four subsets of our dataset: *50-50 gender-balanced sample*, *male speakers*, *female speakers* and the subset of speakers in the previously identified of gender *grey area* (128-148 Hz pitch).

The clustering experiment also provided important insights regarding the accuracy of gender representations of the speaker: none of the obtained clusters showed gender patterns, and none of the best replacements to gender (pitch and jitter) showed relevant mean differences between clusters. Such findings indicate that the most informative representation of the vocal traits of the speaker is not necessarily related with his gender.

Therefore, our final proposal is to move from gender models to one of two representations on the speaker:

1. **Direct replacement to gender** - each speaker would be represented by a *(pitch, jitter)* vector. This representation emulates the distribution obtained by gender labels but now using verifiable and measurable traits of the speaker's voice.
2. **Gender blind representation** - each speaker would be represented by a *(spectral centroid, loudness)* vector. This representation would also be based on verifiable and measurable acoustic features but now providing a depiction of the speaker that is not necessarily related with his gender.

These insights will then be transported to the following research question, by training four different ASR speakers, each containing an uniform distribution over one of following pairs of features: pitch; pitch and jitter; spectral centroid; and spectral centroid + loudness.

5 | RQ2: Measurable balancing in speech applications' train sets

This chapter covers our second research question: *What is the impact (performance and bias) of balancing vocal traits in training datasets for speech applications?*. We explore the hypothesis of using actual vocal representations of the speaker (such as pitch, and loudness) to ensure a more effective and measurable balancing on speech data collections.

To this purpose, we considered the pipeline presented in Figure 5.1. Using a similar architecture, we will train several ASR systems, each with an uniform distribution over a given setup of acoustic features. The only difference between models will be in the distribution of vocal traits in the train set. Hence, we should be able to evaluate the individual impact of a specific setup of vocal traits in the system´s performance and bias. *Post-hoc* analysis on the systems' performance should give us insights on the impact of balancing each of the acoustic features.

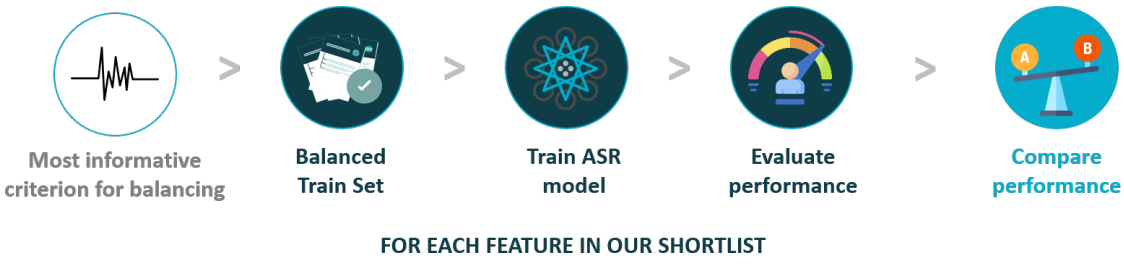

Figure 5.1: High-level experiment pipeline for RQ2.

The remainder of this chapter is structured as follows: Section 5.1 presents the considered train sets, detailing not only the selected vocal traits but also the discretization process of each of the variables; 5.2 presents and discusses the obtained results for the ASR models.

5.1 Selected Vocal Traits for Balancing Speech Data

A key step in our pipeline is the selection and generation of the train sets for the different ASR models. The considered train sets should have an uniform representation for a given setup of variables (hereinafter, *balancement criterion*), i.e, a similar speech time for k fixed-sized groups.

The number of groups considered for balancing (hereinafter, *reference groups*) depends on the variable selected as *balancement criterion*: for categorical variables, the number of groups equals the number of distinct values of the variable; for continuous variables, fixed-size discretization. was

performed. The latter introduces a new variable into our problem: the number of fixed-sized bins to consider in the discretization.

All in all, the selection of the most promising train sets encloses two steps. We start by identifying vocal traits carrying relevant information on the speaker – the most promising *balancement criteria*. Then, for continuous variables, we study the optimal number of bins for each setup of vocal traits. The following subsections detail the obtained results for each of the referred steps. Further details on the considered methodology can be found in Section 3.2.

5.1.1 Criteria for balancing speech data

Considering the insights obtained in Section 4.4, we will explore four acoustic features carrying relevant information on the speaker: pitch (how high and how low a voice is), jitter (pitch fluctuations within the utterance), spectral centroid (a proxy for the brightness of a signal), and loudness (volume). Such variables were, however, divided into two different groups:

1. **Direct replacements to gender** – *pitch* is the most relevant vocal trait to identify the gender of the speaker, being then followed by *jitter*. We will consider two *balancement criteria* on this topic: pitch, and the combination of pitch and jitter.
2. **Gender blind representation** – *spectral centroid* and *loudness* – identified as relevant for differentiating individuals through voice, but not necessarily related with the speaker's gender. From these two variables, spectral centroid was identified as showing the most information on the speaker. Therefore, regarding this group of variables, we will consider two *balancement criteria*: spectral centroid, and the combination of spectral centroid and loudness.

The four features above were analyzed on a speaker level, which implied an aggregation of the features values for all recordings of the same speaker. To this purpose, we considered the aggregation methods identified in Section 4.1: for pitch, spectral centroid, and loudness, median will be used; and for jitter, trimmed mean will be used. These methods were chosen with the purpose of maximizing the inter-speaker variability, i.e., to obtain features with the most differences between speakers.

In addition to the models balanced over vocal traits, we will train a gender-balanced model, and an unbalanced model – i.e., with a random distribution in the training set. These two models, referred to as *Non-Vocal Models*, served as a measure on the impact of balancing datasets over gender labels, and on the impact of maintaining the original distribution of vocal traits, respectively.

All in all, we will consider six different criteria for balancing our train sets: *1) pitch, 2) pitch and jitter, 3) spectral centroid, 4) spectral centroid and loudness, 5) gender and 6) unbalanced/random.*

5.1.2 Optimal number of bins

For train sets balanced over vocal traits (pitch, jitter, spectral centroid, and loudness), discretization will be performed. As a result, the number of bins to consider in the discretization is a relevant variable in our analysis, which ultimately needs to be optimized.

For the purpose of our research, we are mostly interested in train sets with a total size of 200 hours (hereinafter *objective*), and a number of speech hours as similar as possible for all reference groups. Therefore, the maximum number of hours to include in each group is given by the *bin objective* – the ratio between the *objective*, and the number of groups (*k*). Finally, it is important to set naming conventions for the here presented train sets. Since all sets were balanced over a specific combination of vocal traits and bins, each data set will be named with the concatenation of these two elements. As an example, the pitch train set with two reference groups will be referred to as *pitch2*.

We start our study, by setting a minimum of two and a maximum of ten bins for train sets balanced over a single acoustic feature. For train sets balanced over two features (pitch/jitter, and spectral centroid/loudness), the number of bins is given by the product of the number of bins for each of the variables contained in the pair (k_1 and k_2). For this case, we considered bin combinations that assured a minimum of four, and a maximum of twelve bins[1].

Given these constraints, we obtained a total of 34 train sets for vocal models, each with a specific distribution of vocal traits. Then, we focused on ranking the obtained datasets by evaluating their **uniformity** – i.e., if the bins have a similar amount of data (measured by the number of hours of speech recordings). For this purpose, we created the *UniScore*, given by the ratio between the total number of hours in the train set and the *objective* of 200 hours[2]. The score ranges between 0 and 1 – 1 corresponding to a perfectly uniform train set. This measure should capture massive discrepancies in the distribution of hours in the train set, hence detecting the least promising candidates. The scores obtained by each of the 34 train sets are listed in Table 5.1.

As an example, consider an objective of 200 hours, a setup of pitch and jitter as *balancement criterion*, and a total of 12 bins ($k_1 = 3$ and $k_2 = 4$). In a perfect scenario, all 12 bins would fulfill the bin objective, and the total number of hours in the train set would be equal to the objective. For such a scenario, the UniScore would be equal to one, representing the maximum of uniformity.

Such a scenario is, however, quite unusual. For instance, the pitch and jitter train set balanced over 3 and 4 bins, respectively, has a total of 115 hours, and a UniScore of 0.57. To avoid such situations, we defined a minimum of *0.8* for the uniformity score, which reduced our choices to 21 datasets.

Nonetheless, this shortlist may still include train sets with major differences in the distribution of data between reference groups. For instance, the *pitch9* set contains a group that only meets 4.3% of the hours objective, (0.91 hours of a 22.2 *bin objective*). Having such a misrepresentation compromises our objective of obtaining a uniform distribution of data in the dataset.

[1]Possible combinations are the following: 2-2, 2-3, 3-2, 3-3, 3-4, 4-2, 4-3.

[2]Further details on each of the uniformity scores can be found in Section 3.2.

Balancement Criterion	# Bins	Train Size	Bin Objective	Uni Score	BinUni Score	MaxUni Size
Pitch	2	200.00	100.00	1.00	1.00	200.00
Pitch	3	200.00	66.67	1.00	1.00	200.00
Pitch	4	172.95	50.00	0.86	0.46	91.83
Pitch	5	167.47	40.00	0.84	0.19	37.37
Pitch	6	170.08	33.33	0.85	0.10	20.53
Pitch	7	172.69	28.57	0.86	0.04	8.84
Pitch	8	172.96	25.00	0.86	0.04	8.36
Pitch	9	168.48	22.22	0.84	0.04	8.21
Pitch	10	167.47	20.00	0.84	0.04	8.36
Pitch+Jitter	2*2=4	194.98	50.00	0.97	0.90	179.91
Pitch+Jitter	2*3=6	141.86	33.33	0.71	0.13	26.21
Pitch+Jitter	2*4=8	125.24	25.00	0.63	0.04	7.59
Pitch+Jitter	3*2=6	173.61	33.33	0.87	0.21	41.67
Pitch+Jitter	3*3=9	139.04	22.22	0.70	0.02	4.21
Pitch+Jitter	3*4=12	115.44	16.67	0.58	0.00	0.93
Pitch+Jitter	4*2=8	150.64	25.00	0.75	0.11	21.47
Pitch+Jitter	4*3=12	138.05	16.67	0.69	0.00	0.39
Spectral Centroid	2	200.00	100.00	1.00	1.00	200.00
Spectral Centroid	3	200.00	66.67	1.00	1.00	200.00
Spectral Centroid	4	175.01	50.00	0.88	0.64	127.51
Spectral Centroid	5	159.38	40.00	0.80	0.32	64.42
Spectral Centroid	6	159.81	33.33	0.80	0.21	42.37
Spectral Centroid	7	162.29	28.57	0.81	0.14	28.39
Spectral Centroid	8	166.52	25.00	0.83	0.12	24.06
Spectral Centroid	9	163.21	22.22	0.82	0.11	21.90
Spectral Centroid	10	159.38	20.00	0.80	0.10	20.53
Spectral Centroid+Loudness	2*2=4	179.33	50.00	0.90	0.59	117.32
Spectral Centroid+Loudness	2*3=6	155.20	33.33	0.78	0.13	26.69
Spectral Centroid+Loudness	2*4=8	160.48	25.00	0.80	0.06	11.93
Spectral Centroid+Loudness	3*2=6	164.70	33.33	0.82	0.16	32.98
Spectral Centroid+Loudness	3*3=9	155.20	22.22	0.78	0.04	7.13
Spectral Centroid+Loudness	3*4=12	143.06	16.67	0.72	0.02	3.78
Spectral Centroid+Loudness	4*2=8	166.18	25.00	0.83	0.07	13.33
Spectral Centroid+Loudness	4*3=12	146.53	16.67	0.73	0.01	2.04

Table 5.1: Train sets possibilities - UniScore and BinUniScore

To detect such cases, we defined a second filtering criterion: *bin uniformity* – all bins within the train set should have a similar volume of data. To this purpose, we created the *BinUniScore* – the ratio between the minimum number of hours per bin in the train set, and the *bin objective*.

The score once again ranged between 0 and 1 – 1 corresponding to training sets where every group contains a number of hours equivalent to the *bin objective*. Accordingly, we are looking for train sets with a *BinUniScore* as close to 1 as possible, i.e., with a similar number of hours for all reference groups. To this purpose, we once again set a threshold of 0.8, which identified train sets with a maximum difference of 20% in the size of each bin in the dataset.

By the end of this process, we obtained two sets of train sets:

1. **200 hours** - the train sets have a total size around our 200 hours objective (UniScore>0.8) and a maximum difference of 20% in the distribution of hours across bins (BinUniscore>0.8). This set contains five train sets: *pitch* – using 2 and 3 bins –, *pitch* + *jitter* – using a 2-2 combinations –, and *spectral centroid* - using 2 and 3 bins.
2. **100 hours** - the train sets have a total size around our 100 hours objective (UniScore>0.45) and a maximum difference of 20% in the distribution of hours across bins (BinUniscore>0.45). This set contains not only all combinations in the 200 hours group, but also the following combinations: *pitch* – using 4 bins –, *spectral centroid* - using 4 bins –, and *spectral centroid* + *loudness* – using a 2-2 bin combination.

Finally, the previous models will be compared with a gender-balanced and an unbalanced model (hereinafter, *non-vocal models*). Considering that we have two groups of train sets with large differences in their size (100 and 200 hours), we will only compare models in the same group. Therefore, we generated two versions for the train set of the non-vocal models, with a total size of either 100 or 200 hours. By the end of this process, we were left with a total of 17 train sets, as represented in Table 5.2.

5.1.3 How different are the obtained train sets?

Considering that we used distinct criteria for selecting the recordings to include in each of the train sets, it is expected that they show significant differences in the distribution of the vocal traits. Ultimately, if the train sets are significantly different, it is our expectation that the same train sets will also generate ASR models with significant differences in their performance. Therefore, guaranteeing that these differences are significant is crucial to ensure that the conclusions retrieved on the ASRs performance and bias are also significant.

Nonetheless, all train sets were obtained from a similar pool of recordings, which, as thoroughly detailed in Section 3.2, contains 585.97 hours of speech data, from 13,471 different speakers. As a result, in the process of obtaining 17 different samples (either 200 or 100 hours long), it may happen that some of the train sets are not significantly different from others. Indeed, considering that the 200 hours train sets have on average 12,522 different speakers, and that the 100 hours train sets average 10,773 speakers, it is impossible to not repeat speakers between train sets. Indeed, all 200 hours train sets repeat over 90% of their speakers, whilst the 100 hours train sets repeat over 85% of the speakers in their train sets [3].

[3] Further details on the percentage of speakers common to each train set is made available in 7.11, in the Appendix Section

Balancement Criterion	# Bins	# Hours	# Utter.	#Speakers	MeanDur
Gender	2.00	200.00	147,014	12,302	4.90
Gender	2.00	100.00	73,344	10,599	4.91
Pitch	3.00	200.00	148,788	12,730	4.84
Pitch	3.00	200.00	148,113	12,268	4.86
Pitch	2.00	100.00	74,426	11,209	4.84
Pitch	3.00	100.00	73,996	10,709	4.86
Pitch	4.00	97.96	72,536	10,331	4.86
Pitch+Jitter	2*2=4	194.97	146,458	12,575	4.79
Pitch+Jitter	2*2=4	100.00	74,918	10,926	4.81
Random	1.00	200.00	149,266	12,818	4.82
Random	1.00	100.00	74,899	11,251	4.81
Spectral Centroid+Loudness	2*2=4	100.00	76,814	10,368	4.69
Spectral Centroid	2.00	200.00	149,630	12,797	4.81
Spectral Centroid	3.00	200.00	149,403	12,167	4.82
Spectral Centroid	2.00	100.00	74,760	11,306	4.82
Spectral Centroid	3.00	100.00	74,660	10,685	4.82
Spectral Centroid	4.00	100.00	74,419	10,348	4.84

Table 5.2: Shortlist of train sets for ASR.

Nonetheless, each speaker can be represented in our pool of recordings by multiple recordings. Therefore, even if we have a large percentage of repeated speakers between the train sets, they can be represented by a different number of recordings, hence creating differences in the distribution of vocal traits. To test such hypothesis, we computed the number of files per speaker for each of the 17 train sets, and ran a Wilcoxon signed-rank test (α=0.05) for each pair of train sets. P-values over 0.05 will thus indicate us that the pair of train sets not only repeats the speakers, but also their representation in the train set. The obtained p-values for each pair of train sets are presented in Figure 5.2.

This exercise was reproduced separately for the 100 and 200 hours groups, given the discrepancy in the total number of speakers between the 100 and 200 hours train sets. Nonetheless, despite analyzed separately, we will only consider pairs of train sets that are maintained for each of the groups.

Finally, to complement the insights obtained by the previous analysis, we compared the distribution of vocal traits between train sets. To this purpose, for each pair of train sets, the distribution of each acoustic feature (pitch, jitter, spectral centroid and loudness) using a Mann-Whitney U test (α=0.05). This test should thus provides us insights on which vocal traits have a similar distribution, for each pair of train sets[4].

Therefore, the combination of the results from these two tests should not only identify train sets

[4]The obtained results for the 100 and 200 hours train sets are made available in Figures 7.9 and 7.10, in the Appendix Section

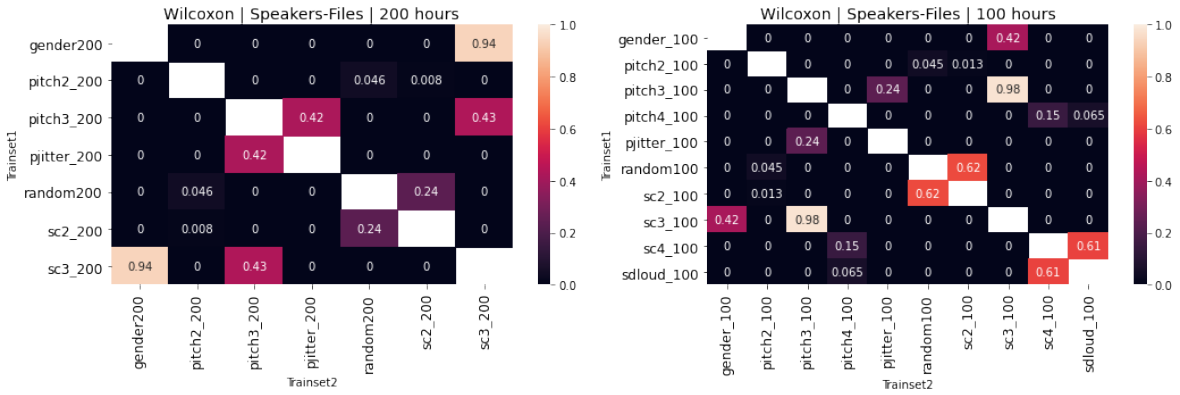

Figure 5.2: Distribution of files per speaker – Wilcoxon Signed Rank test's p-values.

with a similar representation of speakers – hence similar –, but also to identify which vocal traits are mostly responsible for such similarity. Given these conditions, we were able to find three pairs of train sets with a similar representation of speakers:

- **Spectralcentroid2** and the **random** train sets, with 0.24 (200 hours) and 0.62 (100 hours) p-values. — similar both in pitch and jitter.
- **Pitchjitter22** and the **pitch3** train sets, with 0.42 (200 hours) and 0.24 (100 hours) p-values — similar in the spectral centroid distribution.
- **Gender** and **spectralcentroid3**, with 0.94 (200 hours) and 0.42 (100 hours) p-values — similar in the spectral centroid and loudness distribution.

All things considered, the major finding from the list above is the similarity between the spectral centroid 2 and the random train sets, which indicates us that balancing data over two groups of spectral centroid has little impact over the original distribution of vocal traits of the train set.

Moreover, it is worth noting that all pairs of train sets showed significant differences in the distribution of at least one vocal trait, and thus may still lead to significant performance differences when training ASR models over these train sets. For such reason, we will maintain our shortlist of 17 different train sets, and train an ASR model over each of the referred train sets.

5.2 ASR Training

Having identified the most impactful vocal traits for balancing speech data, we will now focus on assessing actual impact of balancing the identified setups of vocal in the train sets for ASR systems. To this purpose, using a common framework, we will train several ASR systems, each with an uniform distribution over one of the four setups of vocal features above – *vocal models*. *Post-hoc* analysis on the systems' performance should give us insights into the impact of vocal traits in the performance of the ASR systems.

For this purpose, we will use the DeepSpeech standard architecture, composed of 5 hidden units (4 ReLU and 1 RNN). The first three are ReLU layers, and the fourth one is an RNN, which includes a set of hidden units with forward recurrence. Finally, the fifth (non-recurrent) layer takes the forward units as inputs. Considering that the focus of our research rests on the vocal traits of the speaker, we did not train any language model and used DeepSpeech's pre-trained language model for all iterations. Further details on the chosen model's parameters can be found in Section 3.2.

Regarding evaluation, Word Error Rate (WER) is our primary performance metric. Considering that our train sets will either have 100 or 200 hours of data, which, as detailed in Section 3.2.5, should guarantee a word error rate (WER) around 45% and 40%, respectively. Given this 5 pp difference in performance, we will only compare the performance of ASR systems trained with a similar amount of data. Instead of obtaining top-of-the-line ASR systems, we will focus on obtaining comparable systems, i.e., trained in a similar context (system architecture, train set size, and training time). Using such a similar setup[5], the only difference between models will be the distribution of vocal traits in the train set, hence we will be able to evaluate the individual impact of a specific setup of vocal traits in system´s performance and bias.

Additionally, it is worth noting that all considered models were trained with a relatively scarce number of hours when compared to state-of-the-art ASR models. Therefore, our analysis will focus on detecting significant performance differences between ASRs, both on the global systems' performance, but also by comparing how biased each model is.

The systems' global performance will be compared using Wilcoxon pairwise tests [40] (α= 0.05), which will evaluate the existence of significant differences in the WER distribution for three different test sets: 1) a 30 hours random sample of the CommonVoice dataset – *cv30*, 2) a 9.5 hours structured sample of the CommonVoice dataset where each speaker is represented by 5 different files – *cvBal*, and 3) the LibriSpeech (other-clean) test set – *libri*. Despite independent from the training data, the two first test sets are expected to have a better performance than the libri set – the recording conditions in which the system was trained are similar to the two first test sets, hence impacting the effectiveness of the acoustic model. Finally, since we have multiple test sets, we used the average of the WER in each test set as our final performance measure.

Bias, on the other hand, will be evaluated by comparing the group performances over two metrics: age (teens, twenties, thirties, forties, fifties, Over60) and gender (male and female) groups. To this

[5]Detailed model architecture can be found in Section 3.2.4

purpose, we will evaluate the existence of significant differences between groups, using either the Wilcoxon pairwise test (α= 0.05 – for gender groups, hence with $k = 2$), or the Kruskal-Wallis test (α= 0.05 – for age groups, hence with $k > 2$). Given that LibriSpeech does not make available the required speaker metadata, our analysis will be limited to the two CommonVoice test sets (30 hours, and 9.5 hours with a balanced distribution).

The following subsections detail the performance results for the aforementioned models. The following subsections detail the obtained performance and bias results for each of the 17 models.

5.2.1 Performance

As stated in Section 5.1.1, we will be working with 17 different train sets (13 vocal models, and 4 non-vocal models), with either 100 or 200 hours of speech data. Given the 100% difference in the size of the train sets, we will only compare models with a similar size. Accordingly, we divided the obtained ASR systems into two groups:

1. **100 hours group**, containing ten different models: pitch (with 2, 3 and 4 reference groups), spectral centroid (with 2, 3 and 4 reference groups), pitch and jitter (with 2-2 bin combination), spectral centroid and loudness (with a 2-2 bin combination), gender and a model with a random distribution.
2. **200 hours group**, containing 7 different models: pitch (with 2, 3 reference groups), spectral centroid (with 2, 3 reference groups), pitch and jitter (with 2-2 bin combination), gender and a model with a random distribution.

Each group is composed not only by models balanced over a given vocal characteristic of the speaker (hereinafter, vocal models), but also by two additional ones: gender and random. The latter two (hereinafter non-vocal models) served as a measure on the impact of balancing datasets over gender labels, and on maintaining the original distribution of vocal traits, respectively.

Performance was measured over three different test sets: *cv30*, *cvBal* and *libri*. Considering that we will calculate the WER for each of the considered test sets, we aggregated the test results by considering an average of the WER. The obtained results are listed in Tables 5.3 and 5.4.

Balancement Criterion	# Hours	WERcv30	WERcvBal	WERlibri	WERmean
Gender	200	44.59%	44.06%	58.67%	49.11%
Random	200	43.01%	42.70%	58.24%	47.98%
Pitch+Jitter 22	200	43.13%	42.76%	57.88%	47.92%
Pitch2	200	43.09%	42.77%	57.38%	47.75%
Pitch3	200	42.63%	42.41%	**56.38%**	47.14%
Spectral Centroid 2	200	45.64%	45.07%	61.11%	50.61%
Spectral Centroid 3	200	**42.12%**	**41.77%**	57.14%	**47.01%**

Table 5.3: Performance results for the 200 hours ASR models

Balancement Criterion	# Hours	WERcv30	WERcvBal	WERlibri	WERmean
Gender	100	50.14%	49.96%	66.16%	55.42%
Random	100	53.19%	52.96%	69.39%	58.51%
Pitch+Jitter 22	100	**48.27%**	**47.85%**	63.35%	**53.16%**
Pitch2	100	50.41%	50.25%	66.04%	55.57%
Pitch3	100	49.91%	49.75%	**65.17%**	54.94%
Pitch4	100	49.44%	49.06%	64.23%	54.24%
SpectralCentroid 2	100	**50.89%**	**50.67%**	64.94%	**55.50%**
SpectralCentroid 3	100	49.51%	49.43%	**62.57%**	53.84%
SpectralCentroid 4	100	49.87%	49.82%	63.64%	54.44%
SpectralCentroid+Loudness	100	48.45%	48.28%	63.27%	53.33%

Table 5.4: Performance results for the 100 hours ASR models

The two CommonVoice test sets show, on average, a 15 pp. lower error rate in comparison to the LibriSpeech set, since these were also obtained from the CommonVoice 6.1 English dataset. Despite independent from the training data, the recording conditions of these two tests are closer to the recording conditions in which the model was trained, hence impacting the effectiveness of the acoustic model. Conversely, performance on the two CommonVoice test sets is similar: 41-46% for the 200 hours models, and 48-53% WER for the models trained with 100 hours of data. Models with 100 hours of data show, on average, a performance 7 p.p. worse than models trained with 200 hours of data.

Further, it is worth noting that no pair of balancement criteria confirmed the null hypothesis of our Wilcoxon pairwise tests [40] (α= 0.05): the median distribution of WER between the two models is similar. Accordingly, no pair of balancement criteria obtained consistent (i.e., for both 100 and 200 hours iterations) and significant p-values results (p-value > 0.05). Hence, all selected balancement criteria lead to significantly different ASR systems.

Vocal models with a minimum of three reference groups are the best performing ones both for the 100 (pitch and jitter, spectral centroid + loudness, pitch 3) and the 200 hours (spectral centroid 3, pitch 3 and pitch and jitter) groups. Indeed, the pattern is clear: three best performing systems are vocal models, with a minimum of three reference groups.

Conversely, non-vocal (gender and random), and pitch 2 show, on average, the worst performances for each of the hour groups. The most extreme example comes from the random model trained with 100 hours, which obtained a 5 point difference in performance when compared to the pitch+jitter model – the top performer in this group.

5.2.1.1 Differences to the Gender-Balanced Model

As noted in Section 2.5.2, data providers focus on balancing speech datasets by assuring uniform distribution gender groups, with the purpose of not only improving performance, but also to prevent the existence bias – i.e., systematic differences in performance between groups of speakers –, when compared to a train set with a random distribution. Thus, the gender model served as baseline for our experiment: our goal is to obtain a performance at least good as the gender model.

To this purpose, the difference in performance to the gender model was calculated (hereinafter, *GenderDiff*). Considering that most of our *balancement criterion*) were tested for the 100 and 200 hours groups, we calculated the mean difference for the two groups. Accordingly, by simultaneously considering the two sets of train sets, this should provide us a final measure to evaluate the global impact of each vocal trait over performance.

Balancement Criterion	**WER**			**GenderDiff**		
	100h	200h	mean	100h	200h	mean
Random	58.51%	47.98%	53.25%	3.09	-1.12	0.99
Pitch2	55.57%	**47.75%**	**51.66%**	0.15	**-1.36**	-0.61
Pitch3	54.94%	47.14%	51.04%	-0.48	-1.97	-1.22
Pitch4*	54.24%	-	54.24%	**-1.18**	-	-1.18
Pitch+Jitter 22	53.16%	47.92%	50.54%	-2.27	-1.19	-1.73
Sdloudness 22*	53.33%	**-**	**53.33%**	-2.09	**-**	-2.09
Spectral Centroid 2	55.50%	50.61%	53.05%	**0.08**	1.5	0.79
Spectral Centroid 3	53.84%	47.01%	50.42%	-1.58	-2.1	-1.84
Spectral Centroid 4*	54.44%	-	54.44%	-0.98	-	-0.98

Table 5.5: Mean Performance Differences to the Gender-Balanced Model

Spectral centroid + loudness (-2.09), spectral centroid 3 (-1.84) and pitch + jitter (-1.73) improved performance the most. On a second level of relevance, we find pitch 3, pitch 4, and spectral centroid 4, which improved performance by 1 pp. Therefore, patterns identified in the previous section were maintained: vocal models with at least 3 reference groups improved, on average, the performance of the gender model in 2 pp.

Conversely, the random model and spectral centroid 2 are the worst performing criteria, obtaining a performance a performance 1 pp greater than the one obtained by gender. Indeed, such results are aligned with the work of Garnerin et al. [5] which suggested that balancing the train set across gender labels improves the performance of speech applications when compared to the original distribution of vocal traits in our pool of recordings.

In addition to this, we see that spectral centroid 2 and the random model show, on average, a similar performance. Such finding is coherent with the insights obtained in Section 5.1.3, where these two train sets were identified not having significant differences, namely on the pitch and jitter distributions.

Finally, vocal models balanced over two reference groups (pitch 2 and spectral centroid 2) showed the most similar performances to gender, with -0.61 and 0.79 differences, respectively. Such pattern is symptomatic of the binary pattern that gender-balanced datasets show: having such a clear division of speakers in two groups reflects on the presence of two major poles in the distribution of vocal traits, each corresponding to a given gender. As identified in Section 4.4 such pattern is more obvious for pitch, which alone is 87.3% accurate in predicting the gender of the speaker.

5.2.2 Bias

Balancing training sets over vocal traits shows thus significant improvements in the global performance of the model. Such improvements could, however, be more significant for specific groups of speakers, hence creating bias. In the context of AI, a biased system is one that shows systematic errors "against" specific sub-groups of people. Conversely, an unbiased system is one that do not show systematic and significant differences in the performance of specific sub-groups.

As stated in Section 3.2, all train sets were obtained by sampling a common pool of recordings. Such pool, however, contains a clear imbalance both in the gender and age representation of speakers: there is a prevalence of male speakers (80-20 ratio), and the 20-30 age range is the dominant one (40%). Such differences may thus provoke significant differences in the performance between groups, hence bias.

Accordingly, it is our interest to analyze if the obtained vocal models are biased towards or against a specific sub-group of individuals. To this purpose, for each of the trained models, we will compare the group performances over two self-reported metrics: age (teens, twenties, thirties, forties, fifties, Over60) and gender (male and female) labels. Given that LibriSpeech does not make available any speaker metadata, our analysis will be limited to the two CommonVoice test sets (30 hours, and 9.5 hours with a balanced distribution).

5.2.2.1 Gender groups

As noted in Section 2.5.2, Garnerin et al. [5] identified bias against women in the performance of speech recognition systems by analyzing the gender representation in four major corpora of French broadcast. The authors concluded that the disparity of available data for both genders caused performance to decrease on women. Accordingly, guaranteeing a similar fair and unbiased systems for gender groups has been a priority not only for the speech industry, but for the AI industry itself.

For the purpose of our experiment, we will consider the binary gender definition (male and female speakers). According to the self-assigned gender labels from the CommonVoice dataset, our original pool of recordings has a 80% males - 20% females distribution, which ultimately might provoke differences in the obtained train sets. To analyze such situation, we will compare the mean performance of each model between gender groups, by calculating, in percentage points, the mean difference in the WER of the two groups. The obtained results are presented

in Table 5.6. Further, we will evaluate the significance of the identified differences using Mann-Whitney U tests (α=0.0.1), hence testing the hypothesis of male and female speakers having a similar WER distribution. The Mann-Whitney test was applied over the concatenation of the two CommonVoice test sets.

Balancement Criterion	**Performance**			**Pvalues**	
	Female	**Male**	**MeanDiff**	**100H**	**200H**
Gender	45.19%	48.04%	-2.848	0.00	0.00
Random	49.02%	48.07%	0.946	0.03	0.32
Pitch3	45.52%	46.25%	-0.731	0.08	0.00
Pitch2	46.36%	46.95%	-0.59	0.00	0.01
Pitch4	49.54%	49.67%	-0.129	0.07	–
Spectralcentroid2	48.33%	48.42%	-0.094	0.01	0.49
Spectralcentroid3	46.56%	46.46%	0.106	0.24	0.06
Spectralcentroid4	50.47%	49.85%	0.614	0.27	–
Sdloudness22	49.71%	48.05%	1.663	0.04	–

Table 5.6: Mean performance per binary gender groups.

As one can see in Table 5.6, only four models reveal the most differences in performance between male and female speakers: the gender, the random model, and the vocal models balanced by the combination of two vocal characteristics: pitch and jitter; and spectral centroid and loudness. There are, however, opposite behaviours within this group of our models: the gender model favours female speakers (3.72 pp lower error rate in comparison to the male group), whereas the other three models favour male speakers (1-2 pp. lower error rate in comparison to the female group).

The Mann-Whitney U test was reproduced both for the 100 and 200 hours iterations of each *balancement criterion*. Accordingly, we will only consider significant the differences that are maintained for iterations of the criteria. Given these constraints, we see that gender (0.00 p-value for both groups), pitch 2 (0.00 and 0.01 p-values for the 100 and 200 hours iterations) and the combination of spectral centroid and loudness (0.04) are the criteria that consistently show significant differences in performance between gender groups.

Conversely, spectral centroid 3 and 4, and pitch 4 are the only criteria that consistently reject the hypothesis of existing significant differences in performance between gender groups. Indeed, models balanced over a single criterion with at least 3 reference groups proved to be the most efficient in preventing performance differences between gender groups.

Differences in the gender model were related with an original gender imbalance of the pool of recordings, which creates a much larger diversity of vocal profiles in the male group. Considering that the gender model was trained with a similar amount of speech data for each of the gender groups, and that the female group shows a smaller diversity of vocal profiles, the obtained model will be optimized for female speakers. Therefore, instead of following the initial behaviour, the gender model seem to favour the minority group. This pattern is clearly identifiable in Figure

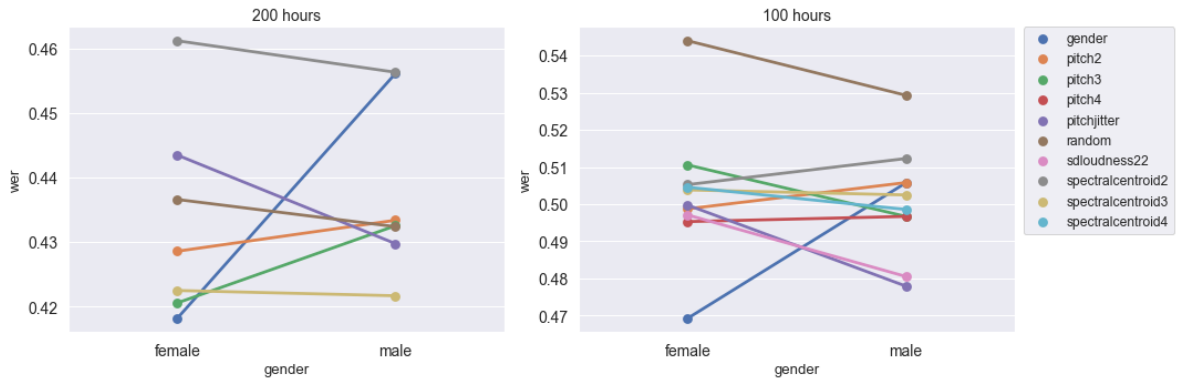

Figure 5.3: Mean performance per binary gender groups.

5.3, where we see that the blue line (corresponding to the gender model), follows the opposite direction to all other models.

Apart from the previous four models, we find that vocal models balanced over a single vocal characteristic have a similar performance across gender groups. Indeed, all models balanced over pitch or spectral centroid seem to be effective in neutralizing performance differences, even if the original pool of recordings is biased towards a specific gender.

5.2.2.2 Age groups

The models' performance across age groups follows a common pattern for all 17 models: speakers in the teens age range (7 pp above the models' average performance) show, on average, a worse performance than all other age groups. Such pattern was reinforced, by conducting a Kruskal-Wallis test which compared the median WER across age groups in each of our test sets. Indeed, all 17 models obtained a p-value under 0.05 for both tests sets, i.e., showed significant performance differences across age groups[6].

Ideally, we are looking for models with a similar performance across each one of these groups, i.e., unbiased both against age and gender groups. Nonetheless, considering the substantial differences in the representation of each of age group in the original pool of recordings, it was not possible to achieve such objective. Therefore, instead of looking for the unbiased models, we focused on identifying the models that were most effective on mitigating performance differences across age groups.

To assess this, for each *balancement criterion*), we estimated the sum of squared differences between the group performance and the model's overall performance (hereinafter *scoreagedifferences*). This should give us an idea on how heterogeneous the model's performance is between groups. The obtained results are presented in the table below.

Pitch 4 and spectral centroid 3 are the two balancement criteria with the lowest score, i.e., with

[6]Detailed p-value results available in Table 7.2 in the Appendix Section

Balancement Criterion	Teens	Thirties	Fourties	Fifties	Over60	Score
Gender	5.66	-3.05	-5.12	-3.45	-7.08	134.1
Random	7.13	-3.12	-4.51	-2.77	-6.14	132.45
Pitch+Jitter 22	7.26	-2.52	-4.66	-1.85	-6.35	131.18
Pitch2	7.79	-2.53	-4.33	-2.55	-5.59	129.51
Pitch3	6.87	-2.38	-3.91	-2.02	-6.24	120.00
Pitch4	6.12	-2.1	-4.74	-1.62	-4.88	96.32
Spectral centroid + loudness (22)	7.09	-2.88	-4.48	-1.33	-5.72	117.76
Spectral centroid2	7.37	-2.67	-4.26	-2.07	-6.37	130.19
Spectral centroid3	8.51	-0.84	-2.35	-0.64	-3.25	105.89
Spectral centroid4	8.47	-2.1	-3.51	-1.9	-5.45	125.54

Table 5.7: Mean performance per age groups, and scoreagedifferences.

the smallest differences in the performance across age groups. On a second level of relevance, we find a group of vocal models with at least 3 reference groups for balancing – spectral centroid loudness, pitch 3, and spectral centroid 4. Indeed, those three models show a difference of about 15 units in the SumSquareDiff score.

On the contrary, the two non-vocal models (random and gender) show the greatest differences between age groups with 134.099 and 132.447 scores, respectively. Differences in the gender model were once again related with an original imbalance of the pool of recordings, specifically concerning the distribution of genders for each of the age groups.

Finally, it is worth noting the results obtained by the pitch 2, spectral centroid 2 and the pitch + jitter models, which obtained a score very close to the obtained by the two non-vocal models. Accordingly, similarly to the insights obtained in the performance section, the vocal models balanced with 2 groups seem to not have a behaviour too different from the obtained by the gender model.

Gathering these insights, we conclude that vocal models with a minimum of three reference groups have the least differences in performance across age groups, hence being the most effective setups of vocal traits in mitigating age bias.

5.3 Most impactful acoustic features

This chapter focused on answering our second research question: *What is the impact (performance and bias) of balancing vocal traits in training datasets for speech applications?*. Our hypothesis is that using actual vocal representations of the speaker (such as pitch, and loudness) to drive the speech data collections offers a more effective solution than prevalent methods that based on self-reported gender labels.

Results show that vocal models with a minimum of three reference groups show a 1-2 pp. significant improvement in performance when compared to the gender-balanced model. Conversely, the random model and spectral centroid 2 were the worst performing models, obtaining a performance one point greater than the gender-balanced model. Such differences are indeed coherent with our baseline hypothesis that balancing the train set across gender improves the performance of the ASR systems. Further, as stated by Garnerin et al. [5], we see that the gender-balanced model still produces better performance results than systems trained over random distributions of vocal traits.

Vocal models balanced over two reference groups (pitch 2 and spectral centroid 2) showed the most similar performances to gender. Such pattern is symptomatic of the binary pattern that gender-balanced datasets show: having such a clear division of speakers in two groups reflects on the presence of two major poles in the distribution of vocal traits, each corresponding to a given gender. As discussed in Section 4.4 such pattern is more obvious for pitch, which alone is 87.3% accurate in predicting the gender of the speaker.

Balancing training sets over vocal traits shows thus significant improvements in the global performance of the model. Such improvements could, however, be more significant for specific groups of speakers, hence creating bias. To this purpose, for each of the trained models, we analyzed the group performances for two self-reported metrics: age (teens, twenties, thirties, forties, fifties, Over60) and gender (male and female) labels.

Concerning gender, only four balancement criteria reveal significant differences in performance between male and female speakers: the gender, the random model, and the two vocal models balanced by two vocal traits (pitch and jitter; and spectral centroid and loudness). There are, however, opposite behaviors within this group of our models: the gender-balanced model favors female speakers, whereas the other three models favor male speakers (1-2 pp. lower error rate in comparison to the female group). Differences in the gender-balanced model were related with an original imbalance of the pool of recordings (80% males – 20% females), which creates a much larger diversity of vocal profiles in the male group.

On the other hand, results show that vocal models balanced over a single vocal characteristic have a similar performance across gender groups. Indeed, all models balanced over pitch or spectral centroid seem to be effective in neutralizing performance differences, even if the original pool of recordings is biased towards a specific gender.

Regarding age groups, vocal models with a minimum of three groups showed the least differences in performance across age groups, hence contributing to mitigate prevailing bias in the input data.

On the contrary, the two non-vocal models (random and gender) show the greatest differences between age groups. Differences in the gender-balanced model were once again related with an original imbalance of the pool of recordings, specifically concerning the distribution of genders for each of the age groups.

Overall, jointly considering the two evaluation dimensions (performance and bias), the best performing models were the ones balanced over a single vocal characteristic of the speaker (pitch or spectral centroid), with a minimum of three reference groups. Indeed, all models balanced over pitch or spectral centroid seem to be effective in neutralizing performance differences, even if the original pool of recordings is biased towards a specific gender. Within this group, the setup composed by spectral centroid and three reference groups (spectral centroid 3) consistently showed the most improvement over performance and bias.

6 | Discussion

As algorithms drive more decision-making processes, machine learning models' tendency to learn our input data biases is a massive problem. Furthermore, the wide range of new diverse, and heterogeneous users demands robust and unbiased solutions that perform successfully regardless of their individual characteristics or demographics.

In the specific case of speech applications, research identified systematic errors against social groups of our society, such as female speakers, elderly speakers, or even misrepresented ethnic groups. To fight this, data providers' most prevalent interventions focus on assuring uniform distributions over the same aspects in which bias is detected, particularly across binary gender groups. However, balancing data along these features has three major drawbacks. First and foremost, as detailed in Section 2.4, these features are hard to test against when collecting audio for training such systems (particularly in a remote collection scenario). Secondly, they do not represent the individual's actual vocal traits (being only proxies of that). Finally, if used incorrectly, these proxies can be dangerous in the sense that they may be perpetuating social stereotypes (for instance, what a male voice is expected to sound like).

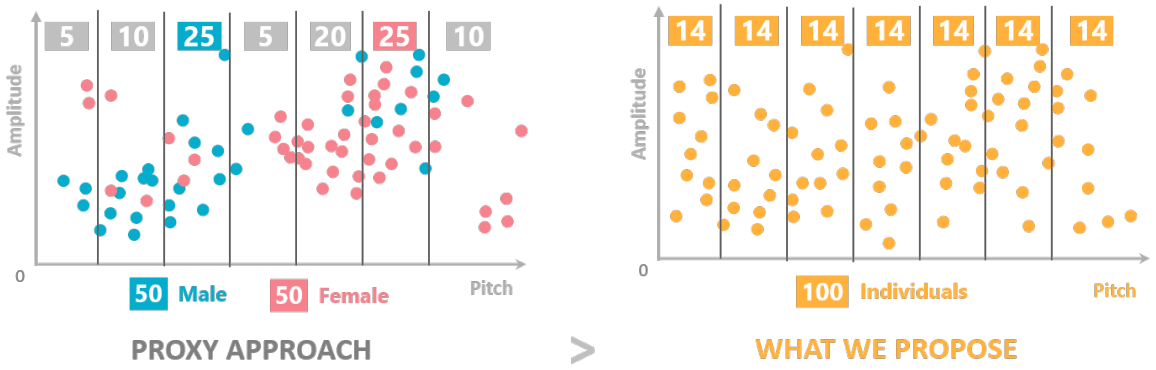

Figure 6.1: Research hypothesis: moving from proxy to actual vocal traits as balancement criterion.

Given this background, this work explored the hypothesis of replacing gender proxies with actual vocal representations of the speaker to drive the data collection process. Recalling the Figure presented in Chapter 3, we can think of a hypothetical situation where a speech dataset containing 100 speakers is adjusted to ensure a 50-50 binary gender distribution. Figure 6.1 replicates this scenario by mapping the speaker distribution in the instance space using two vocal traits: pitch and amplitude. In an ideal scenario, this intervention would guarantee a similar representation of voice profiles in the dataset. The actual scenario is, however, a lot different: despite existing a dominant vocal profile within each gender group, the distribution of the vocal traits is not

homogeneous. As one can see, there is a broader diversity spectrum for which gender proxies are relatively short in representing, which ultimately leads to a misrepresentation of speakers that fail to follow the typical vocal profile of their gender group. In addition, as stated in Section 2.4, the considered gender stats are quite hard to verify and contest in a crowdsourcing context.

Hence, our hypothesis is that using actual vocal representations of the speaker (such as pitch, and loudness) to drive the speech data collections offers a more effective solution (performance and bias) than prevalent methods that are based on self-reported gender labels. To test this, we divided our research in two major phases. First, we identified a shortlist of acoustic features (representing vocal traits) that are capable of characterizing and identifying individuals through voice. Then, we evaluated the impact (performance and bias) of balancing such features in training datasets for speech applications. Thus, our baseline objective was to identify a uniform distribution of vocal traits that can at least ensure a similar performance to a speech application trained over a dataset with a 50-50 gender distribution (hereinafter, gender-balanced models). Even if we obtain the same results, vocals should offer a more objective, hence verifiable method for the collection process than the gender-balanced model.

The identification of relevant acoustic features was the first milestone in our study. From the Sharma et al. [15] taxonomy described in Section 2.2, we explored the subset of acoustic features that meet two conditions: 1) verifiable and 2) semantically understandable. In addition to this, we focused on identifying features that are capable of ensuring diversity in the train set, i.e., that show high variability in the instance space. Once these conditions were applied, we obtained an initial list of nine acoustic features: spectral centroid, spectral spread, HNR, pitch, jitter, shimmer, loudness, energy and speaking rate.

All features were extracted on a utterance level, i.e., for each audio file in our dataset. Features on this level, however, revealed an extremely high variability – an expected pattern considering that the files have a five seconds average duration. Indeed, acoustic features computed over files with such a small length typically show a great fluctuation, hence affecting the significance of the obtained measures. To overcome this limitation, we aggregated the features on a speaker-level, i.e., aggregated the features values for all recordings of the same speaker. To this purpose, we imposed a minimum of 20 seconds of recordings per speaker, corresponding to an average of 4 entries per speaker. Such threshold may, however, be optimized in future work. Indeed, increasing the total duration per speaker may lead to more stable and significant features, and ultimately provide an even more objective of the speaker.

Having defined our pool of acoustic features, we divided our investigation for the most informative vocal traits into two intermediate milestone: 1) *direct replacements to gender* – identify features that mimic the distribution of vocal traits between gender groups, – and 2) *gender blind* – identify features not necessarily related with gender, but capable of distinguishing speakers through voice. The reason for this division was once again aligned with our baseline objective of obtaining a similar performance and bias results of a model trained with a 50-50 gender distribution. Accordingly, our expectation was that the features in the *direct replacement to gender* group ensure a performance and bias impact similar to the one obtained to gender labels, and that the *gender blind* features offer a blind, hence independent representation of the vocal traits. To this purpose, two

major experiments were conducted: 1) gender classification via traits, 2) building clusters over the speaker-aggregated features, and identify the acoustic features explaining most differences between clusters.

Results show that pitch and jitter emulate 89.1% of the speaker information provided by gender labels, and that pitch alone conveys 87.3% of the gender information. Pitch, indeed, is quite efficient in predicting the gender in our data except for speakers located in the 128-148 Hz pitch range. For such interval, jitter provides important information to separate genders by mapping males to higher jitter values. Such results are aligned with the work of Singh [13] which states that the adult woman's average pitch range is from 165 to 255 Hz, while a man's is 85 to 155 Hz.

Gender, however, may not be the most accurate representation of the speaker. Accordingly, our second group of features – *gender blind representations* – contains vocal traits not necessarily related with gender but still conveying valuable information to differentiate speakers. The obtained results indicate spectral centroid and loudness as the key variables discriminate vocal profiles - both weakly correlated with gender. Indeed, when building clusters of speakers, the two explained most of the differences between the obtained groups. This finding was confirmed over four subsets of our dataset: *50-50 gender-balanced sample*, *male speakers*, *female speakers* and the subset of speakers in the previously identified of gender *grey area* (128-148 Hz pitch). Accordingly, spectral centroid and loudness revealed as the most relevant features for separating and differentiating individuals through voice.

When comparing the obtained clusters with gender groups, we see that none of the identified gender replacements (pitch, jitter) showed significant differences between clusters. Further, no gender pattern was identified in the obtained, even for a two cluster partition which ultimately should correspond to a gender partition of the individuals. Instead of showing a clear separation between male and female speakers, the obtained clusters were consistently separated by two acoustic features independent from gender: spectral centroid and loudness. These findings were not only maintained for different granularity levels (subsets of speakers with different dimensions and profiles, and for different partition levels), but also for each of the gender groups. Therefore, these results reinforce our hypothesis that gender labels are not the optimal representation of the speaker, offering a narrow depiction of the individual. Ultimately, that a collection driven by vocal traits like spectral centroid and loudness would be more diverse than one based on gender labels.

By the end of this analysis, we obtained a shortlist of four setups of vocal features: pitch, pitch + jitter – emulating the distribution of vocal traits obtained by gender labels –, and spectral centroid and spectral centroid + loudness – the most effective and important acoustic features to differentiate speakers through voice.

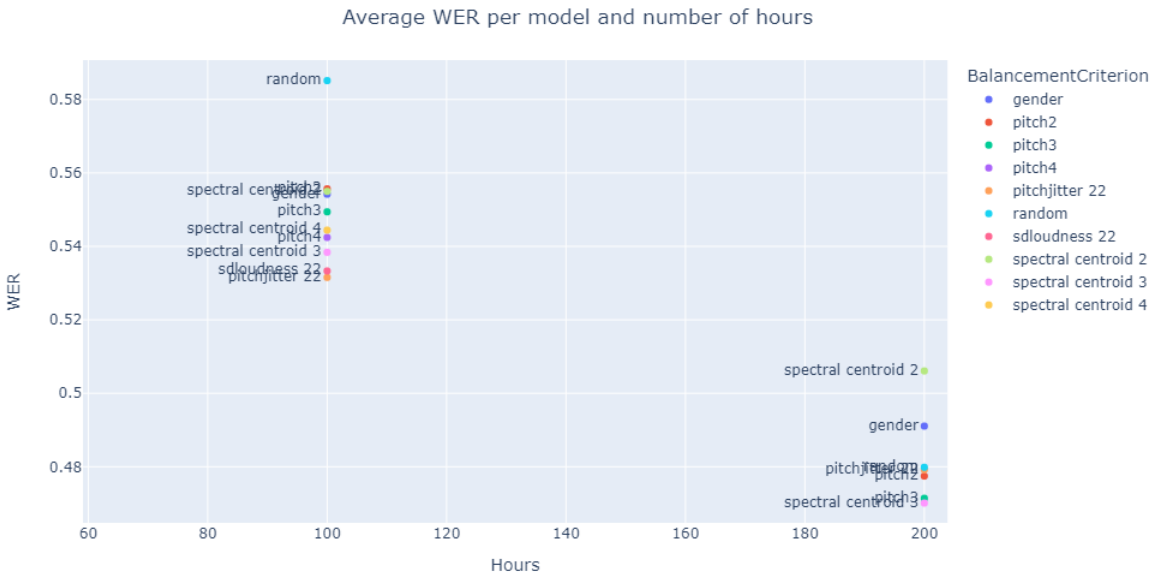

Figure 6.2: Mean performance per balancement criterion and train set size

Having defined our shortlist of acoustic features, we evaluated the impact of balancing such variables in the train sets for speech applications. To this purpose, using a common framework, we will train several ASR systems, each with an uniform distribution over one of the four setups of vocal features above – *vocal models*. Using such a similar setup, the only difference between vocal models will be the distribution of vocal traits in the train set, hence we will be able to evaluate the individual impact of a specific setup of vocal traits in system´s performance and bias. Additionally, to obtain a performance and bias baseline, two more models were trained: a gender-balanced model, and an unbalanced model – i.e., with a random distribution in the training set. These two models, referred to as *Non-Vocal Models*, served as a measure on the impact of balancing datasets over gender labels, and on the impact of maintaining the original distribution of vocal traits, respectively.

A crucial step in our experiment pipeline is the selection and generation of the train sets for the vocal models. The considered train sets should have an uniform representation for the selected setup of features (hereinafter, *balancement criterion*), i.e, a similar speech time for k fixed-sized groups (hereinafter, *reference groups*). Considering that all acoustic features are continuously-valued, the definition of our reference groups was made by applying fixed-sized discretization, using a variable number of reference groups. As a result, the number of reference groups to consider in the discretization is a relevant variable in our analysis, ultimately compromising the uniformity of the obtained train sets. Given this context, uniformity requirements were introduced in the train sets, having we obtained a shortlist of 17 train sets, with either 100 or 200 hours of speech data, and a maximum of 4 reference groups. Finally, since all sets were balanced over a specific combination of vocal traits and bins, each data set was be named with the concatenation of these two elements. As an example, the pitch train set with 2 reference groups was be referred to as *pitch2*.

Results show that vocal models with a minimum of three reference groups show a 1-2 pp. significant improvement in performance when compared to the gender-balanced model. Conversely, the random model and spectral centroid 2 were the worst performing models, obtaining a performance one point greater than the gender-balanced model. Such differences are indeed coherent with our baseline hypothesis that balancing the train set across gender improves the performance of the ASR systems. Further, as stated in 2.5.2, we see that the gender-balanced model still produces better performance results than systems trained over random distributions of vocal traits.

Vocal models balanced over two reference groups (pitch 2 and spectral centroid 2) showed the most similar performances to gender. Such pattern is symptomatic of the binary pattern that gender-balanced datasets show: having such a clear division of speakers in two groups reflects on the presence of two major poles in the distribution of vocal traits, each corresponding to a given gender. As discussed in Section 4.4 such pattern is more obvious for pitch, which alone is 87.3% accurate in predicting the gender of the speaker.

Balancing training sets over vocal traits shows thus significant improvements in the global performance of the model. Such improvements could, however, be more significant for specific groups of speakers, hence creating bias. To this purpose, for each of the trained models, we analyzed the group performances for two self-reported metrics: age (teens, twenties, thirties, forties, fifties, Over60) and gender (male and female) labels.

Concerning gender, only four balancement criteria reveal significant differences in performance between male and female speakers: the gender, the random model, and the two vocal models balanced by two vocal traits (pitch and jitter; and spectral centroid and loudness). There are, however, opposite behaviors within this group of our models: the gender-balanced model favors female speakers, whereas the other three models favor male speakers (1-2 pp. lower error rate in comparison to the female group). Differences in the gender-balanced model were related with an original imbalance of the pool of recordings (80% males – 20% females), which creates a much larger diversity of vocal profiles in the male group.

On the other hand, results show that vocal models balanced over a single vocal characteristic have a similar performance across gender groups. Indeed, all models balanced over pitch or spectral centroid seem to be effective in neutralizing performance differences, even if the original pool of recordings is biased towards a specific gender.

Regarding age groups, vocal models with a minimum of three groups showed the least differences in performance across age groups, hence contributing to mitigate prevailing bias in the input data. On the contrary, the two non-vocal models (random and gender) show the greatest differences between age groups. Differences in the gender-balanced model were once again related with an original imbalance of the pool of recordings, specifically concerning the distribution of genders for each of the age groups.

Given this background, both evaluation dimensions (performance and bias) confirm our initial hypothesis that using actual vocal representations of the speaker (such as pitch, and loudness) to drive the speech data collections offers a more effective solution (performance and bias) than prevalent methods that are based on self-reported gender labels. Models balanced over a single

vocal characteristic of the speaker (pitch or spectral centroid), with a minimum of three reference groups. Indeed, all models balanced over pitch or spectral centroid seem to be effective in neutralizing performance differences, even if the original pool of recordings is biased towards a specific gender. Within this group, the setup composed by spectral centroid and three reference groups (spectral centroid 3) showed the most improvement in performance and bias.

Vocal models balanced over two reference groups (pitch 2 and spectral centroid 2) showed the most similar performances to gender. Such pattern is symptomatic of the binary pattern that gender-balanced datasets show: having such a clear division of speakers in two groups reflects on the presence of two major poles in the distribution of vocal traits, each corresponding to a given gender. As discussed in Section 4.4 such pattern is more obvious for pitch, which alone is 87.3% accurate in predicting the gender of the speaker, hence validating the insights obtained in RQ1.

Finally, as stated in section 1, we do not deny that measuring systems' performance across social groups (like the ones provided by gender information) is still relevant. Indeed, our results are aligned with the work of Garnerin et al. [5], which recommends a similar representation of genders in the train sets for speech applications: when compared to the random model, the gender-balanced model showed a better performance and bias results. However, this kind of sensitive self-reported metadata must not be contested on the basis of normative (and potentially offensive) approaches, and for that reason, are not fit to drive data collection.

When compared to the prevalent method based on self-reported gender labels, the here identified vocal traits (particularly pitch and spectral centroid) offer a more verifiable, ethical and effective approach to collection of speech data. Verifiable since they are measurable and objective depictions of the speakers instead of self-reported and hardly verifiable labels. Effective since they improved performance and reduced bias across gender and age groups. Ethical in the sense that they actual and fact-based depictions of the individual, independent of its ethnicity, age, gender, etc.

Accordingly, our hypothesis that using actual vocal representations of the speaker (such as pitch, and loudness) to drive the speech data collections offers a more effective solution (performance and bias) than prevalent methods that are based on self-reported gender labels was verified for train sets balanced with a single vocal train set, and at least two reference groups. Indeed, the proposed technique guaranteed a 1-2 pp. reduction in the models' error rate, and proved to be effective in mitigating bias conflicts that were identified in the gender-balanced model. Pitch and spectral centroid, calculated on a speaker-level, proved to be the most impactful vocal traits, with a specific highlight on spectral centroid which not only obtained the best performance, but also produced the most unbiased models.

7 | Conclusion and Future Work

As the machine learning industry expands to more natural forms of interaction with everyday devices and services, such as communication via natural language, the need for robust and unbiased solutions that perform successfully regardless of their individual characteristics or demographics.

In the specific case of speech applications, literature unveiled systematic errors against social groups of our society, such as female speakers [5], elderly speakers [41, 9], or even misrepresented ethnic groups [10]. As an answer to these conflicts, data providers have been focusing on balancing speech datasets by assuring uniform distributions over binary gender groups. Such interventions, however, pose three major limitations. First and foremost, gender is only a proxy for actual vocal characteristics of the speaker, and for that reason, may not represent the complete spectrum of speaker diversity. Secondly, as previously referred, these interventions are based on self-reported data, which is hard to contest. Finally, limiting speaker classification to gender labels perpetuates social stereotypes, for instance, of what a male voice is expected to sound like.

Given this context, this work explored the hypothesis of replacing gender proxies with actual vocal representations of the speaker to drive the data collection process. Our hypothesis is that using actual vocal representations of the speaker (such as pitch, and loudness) to drive the speech data collections offers a more effective solution (performance and bias) than prevalent methods that are based on self-reported gender labels.

To this purpose, our research was driven using a concrete speech application (an automatic speech recognizer), concluding on which vocal traits should be uniformly represented in the training dataset and measuring the impacts of such distribution in the model's performance and biases. Briefly, we considered the following two research questions: 1) Which voice traits better differentiate and characterize speakers?, and 2) What is the impact of balancing such features in the training dataset of a speech application? Our contributions are the following:

1. **Which voice traits better differentiate and characterize speakers?**
   Four setups of acoustic features were identified as effective and measurable descriptors of the speaker: pitch, pitch + jitter – emulating the distribution of vocal traits obtained by gender labels –, and spectral centroid and spectral centroid + loudness – the most effective and important acoustic features to differentiate speakers through voice.
2. **What is the impact (performance and bias) of balancing vocal traits in training datasets for speech applications?**
   Results show that balancing vocal traits with at least three reference groups leads to significant improvements in performance of the models and mitigates bias conflicts in the original model. In comparison to a model balanced over gender labels, performance improved by two percentage points, and bias was reduced both across age and gender groups. Further, considering that prevailing balancing interventions are based on self-reported speaker

metadata (such as gender and age), this method offers a more effective and variable method to drive the collection process of speech data.

Overall, jointly considering the two evaluation dimensions (performance and bias), the best performing models were the ones balanced over a single vocal characteristic of the speaker (pitch or spectral centroid), with a minimum of three reference groups. Indeed, all models balanced over pitch or spectral centroid seem to be effective in neutralizing performance differences, even if the original pool of recordings is biased towards a specific gender. Within this group, the setup composed by spectral centroid and three reference groups (spectral centroid 3) consistently showed the most improvement over performance and bias.

Vocal models balanced over two reference groups (pitch 2 and spectral centroid 2) showed the most similar performances to gender. Such pattern is symptomatic of the binary pattern that gender-balanced datasets show: having such a clear division of speakers in two groups reflects on the presence of two major poles in the distribution of vocal traits, each corresponding to a given gender. As discussed in Section 4.4 such pattern is more obvious for pitch, which alone is 87.3% accurate in predicting the gender of the speaker.

When compared to the prevalent method based on self-reported gender labels, the here identified vocal traits (particularly pitch and spectral centroid) offer a more verifiable, effective and ethical approach to the collection of speech data: verifiable since they are measurable and objective depictions of the speaker; effective since they improve performance by two percentage points and reduce bias both across gender and age groups; and ethical in the sense that they are actual and fact-based representations, blind to the speaker´s ethnicity, age, gender, etc.

Also worth noting is that the concept of gender is a subjective, hence evolving one. If we consider that the self-reported gender labels offer a binary representation of the speaker (male or female), the usage of gender labels to drive the collection process is itself discriminatory against all individuals who identify with non-binary gender identities. An approach based on vocal traits would be blind to all social groups, hence ensuring that speakers are represented as an individual element, and not by the typical voice of their social group.

In a crowdsourcing context, this approach could not only prevent quality issues in the collected data (eg. a misalignment between the self-reported gender labels, and the actual gender of the speaker), but also represent important savings, by reducing the number of validation tasks introduced in the collection pipeline.

The results of this study, while interesting, should not be taken as generalizations for all individuals. Further research is needed in order to assess if these findings are maintained for other languages, nationalities and recording conditions (which may, for instance, impact the loudness of the recording). Future work also includes testing alternative setups for the train sets, namely by considering a different distribution (eg. *playkurtic* distribution, or the normal distribution) for the balancing process. Additionally, it is worth noting that the usage of a larger and more diverse pool of recordings may allow a greater number of reference groups for balancing. Ultimately, the obtained train sets would not only be larger (hence more accurate), but also more balanced – potentially leading to even more significant impacts over the systems' performance and bias.

Appendix

A Research Question 1

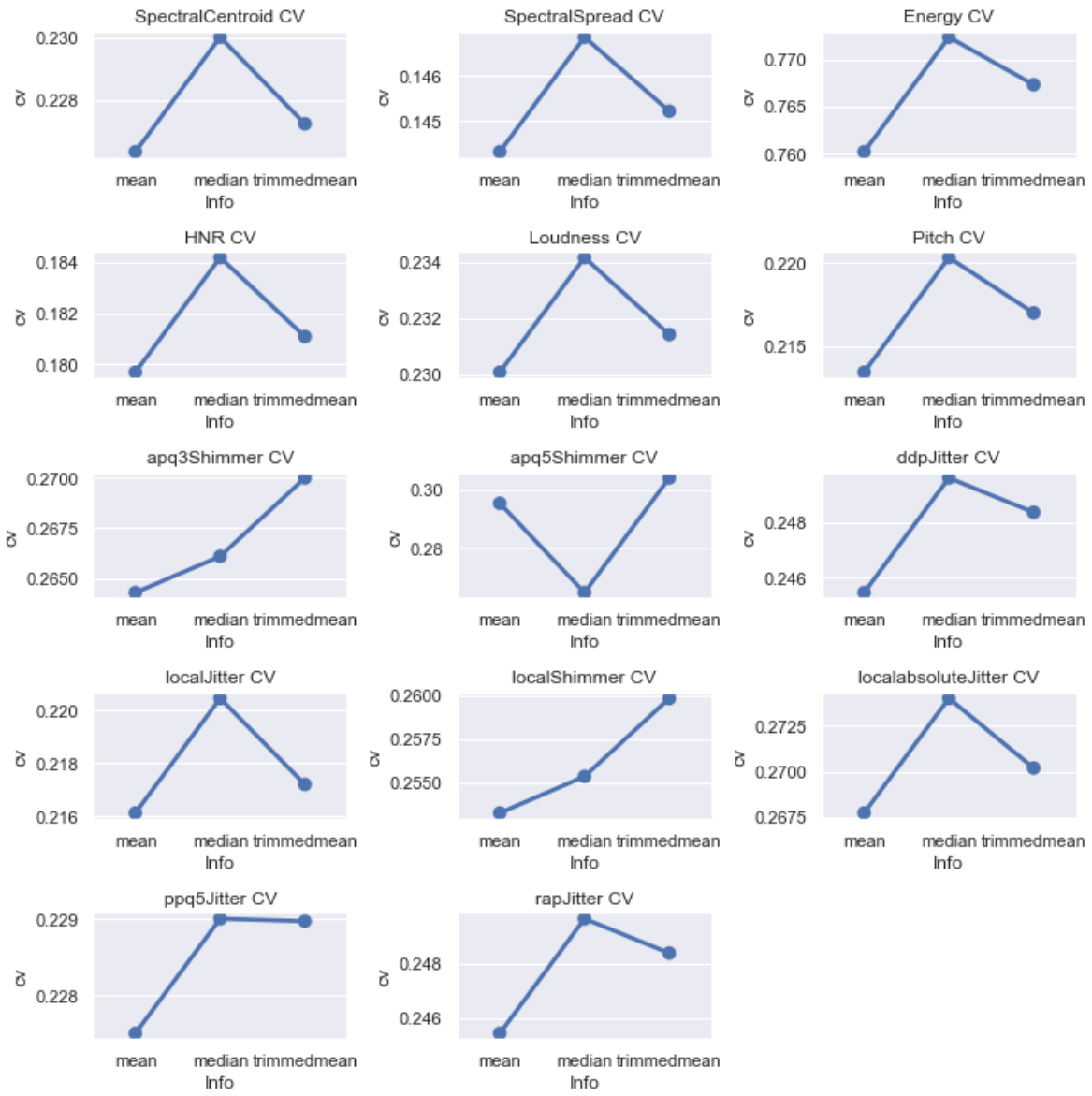

Figure 7.1: Inter-speaker variability per variable, for each aggregation method.

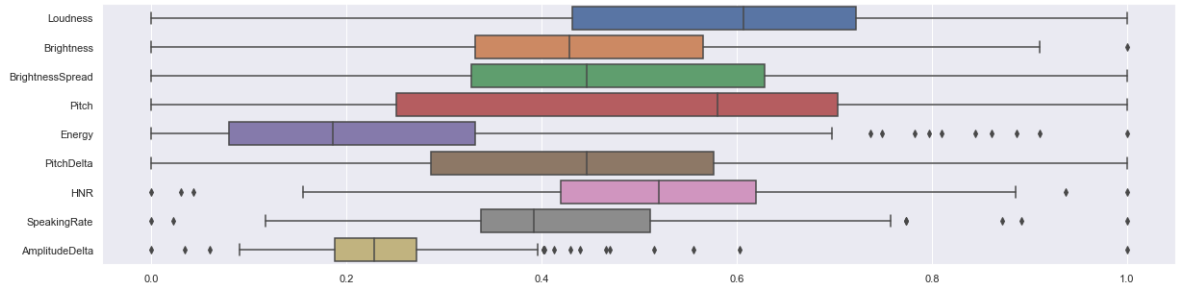

Figure 7.2: Box-plots for each acoustic feature in our RQ1 shortlist.

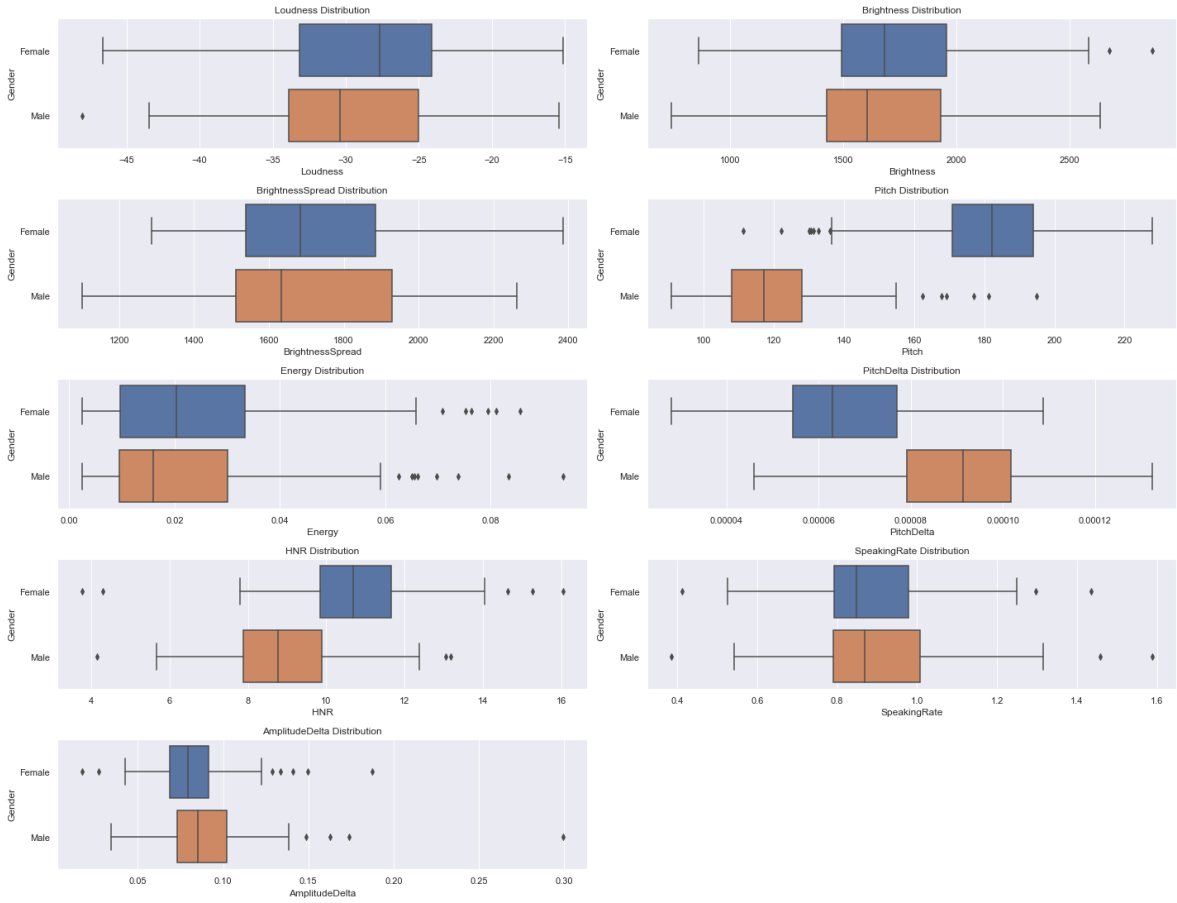

Figure 7.3: Gender mean differences for each acoustic feature in the RQ1 shortlist.

Pitch range	Female	Male	Fabs	CumFemale	CumMale	CumInterval	%IntervalF	%IntervalM	%CumF	%CumM
228-223	2	0	2	2	0	2	100	0	100	0
223-218	1	0	1	3	0	3	100	0	100	0
218-213	6	0	6	9	0	9	100	0	100	0
213-208	8	0	8	17	0	17	100	0	100	0
208-203	4	0	4	21	0	21	100	0	100	0
203-198	9	0	9	30	0	30	100	0	100	0
198-193	19	1	20	49	1	50	95	5	98	2
193-188	16	0	16	65	1	66	100	0	98	2
188-183	21	0	21	86	1	87	100	0	99	1
183-178	18	1	19	104	2	106	95	5	98	2
178-173	20	1	21	124	3	127	95	5	98	2
173-168	17	1	18	141	4	145	94	6	97	3
168-163	5	1	6	146	5	151	83	17	97	3
163-158	3	1	4	149	6	155	75	25	96	4
158-153	9	2	11	158	8	166	82	18	95	5
153-148	5	2	7	163	10	173	71	29	94	6
148-143	4	5	9	167	15	182	44	56	92	8
143-138	1	2	3	168	17	185	33	67	91	9
138-133	2	3	5	170	20	190	40	60	89	11
133-128	5	5	10	175	25	200	50	50	88	12
128-123	0	10	10	175	35	210	0	100	83	17
123-118	1	13	14	176	48	224	7	93	79	21
118-113	0	15	15	176	63	239	0	100	74	26
113-108	1	10	11	177	73	250	9	91	71	29
108-103	0	12	12	177	85	262	0	100	68	32
103-98	0	5	5	177	90	267	0	100	66	34
98-93	0	5	5	177	95	272	0	100	65	35
93-88	0	3	3	177	98	275	0	100	64	36

Table 7.1: Male and female distribution for 5Hz fixed-sized pitch intervals.

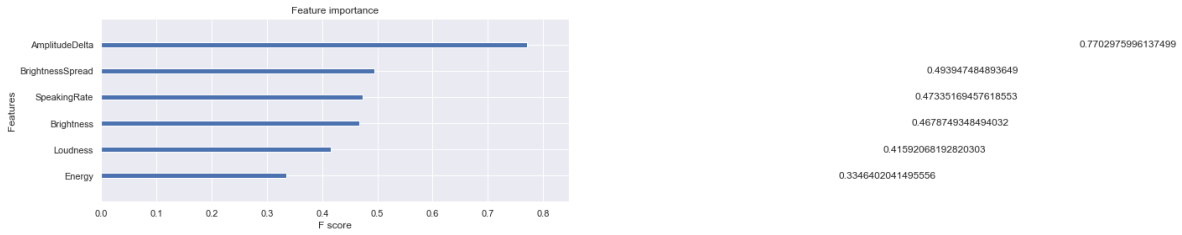

Figure 7.4: Feature contribution for the model excluding acoustic features highly correlated with gender.

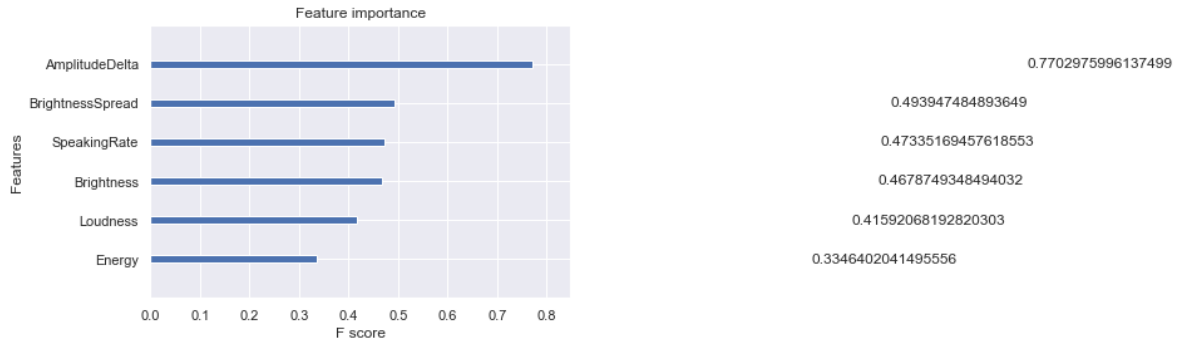

Figure 7.5: Feature contribution for the model excluding pitch information.

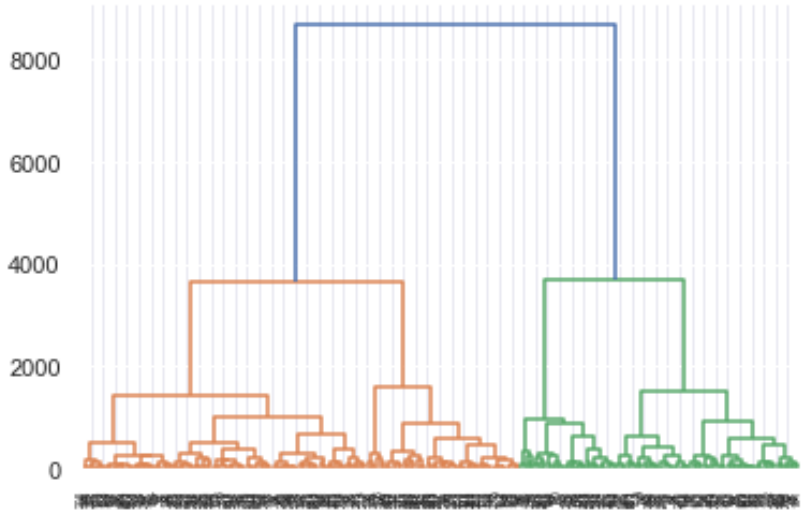

Figure 7.6: Dendogram for the complete dataset.

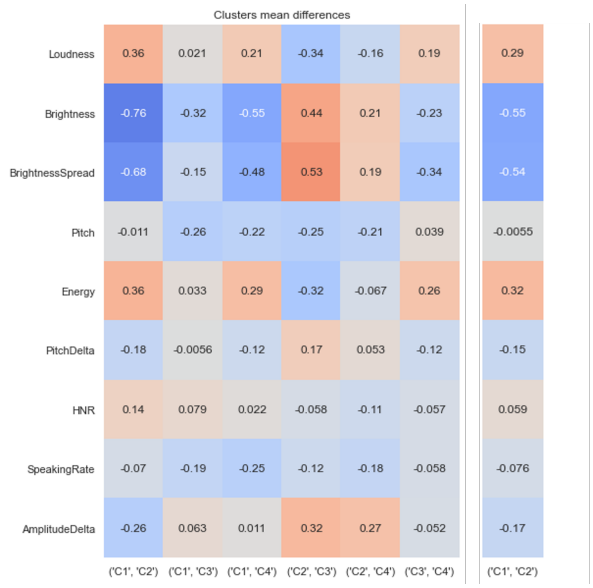

Figure 7.7: Standardized mean cluster differences for speakers in the 130-150 Hz pitch range.

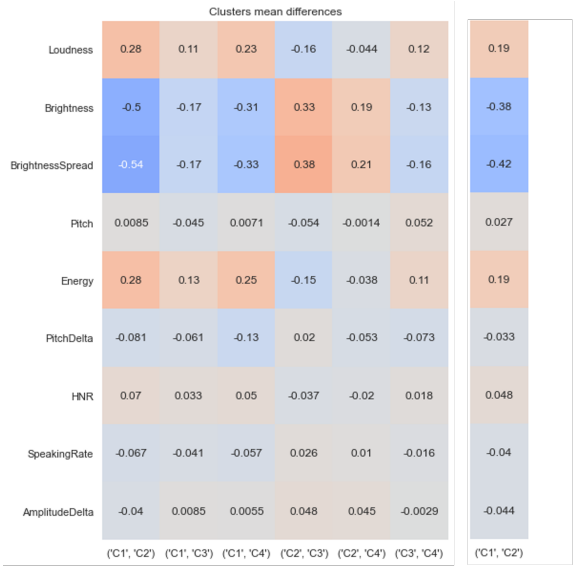

Figure 7.8: Standardized mean cluster differences for the female subset of speakers.

B Research Question 2

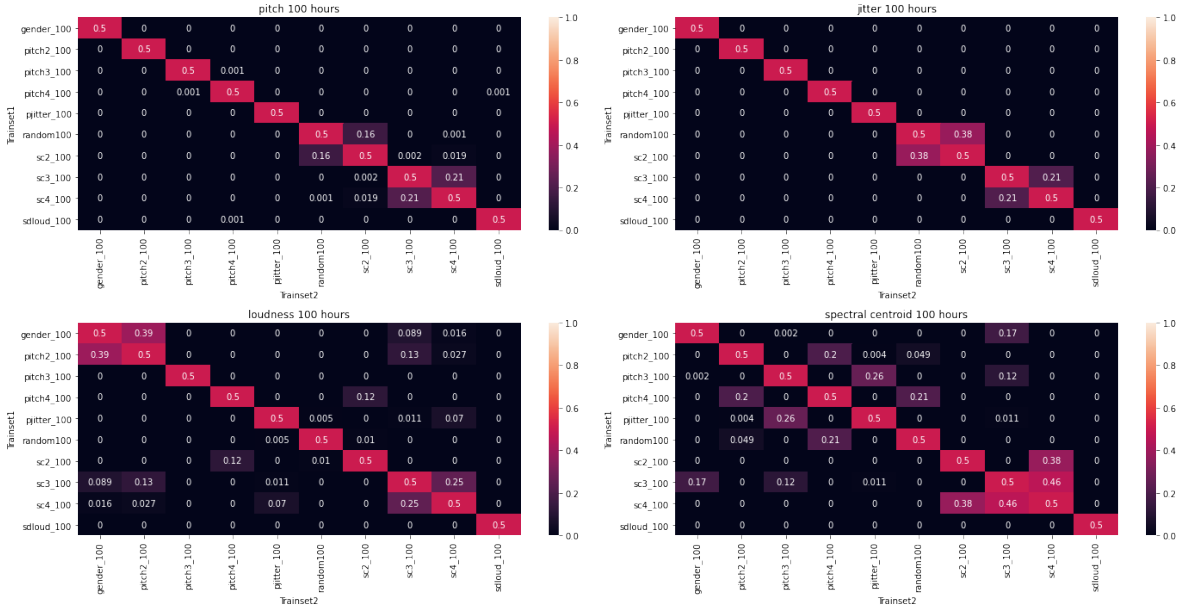

Figure 7.9: Mann-Whitney U p-values for each pair of 100 hours train sets and acoustic feature.

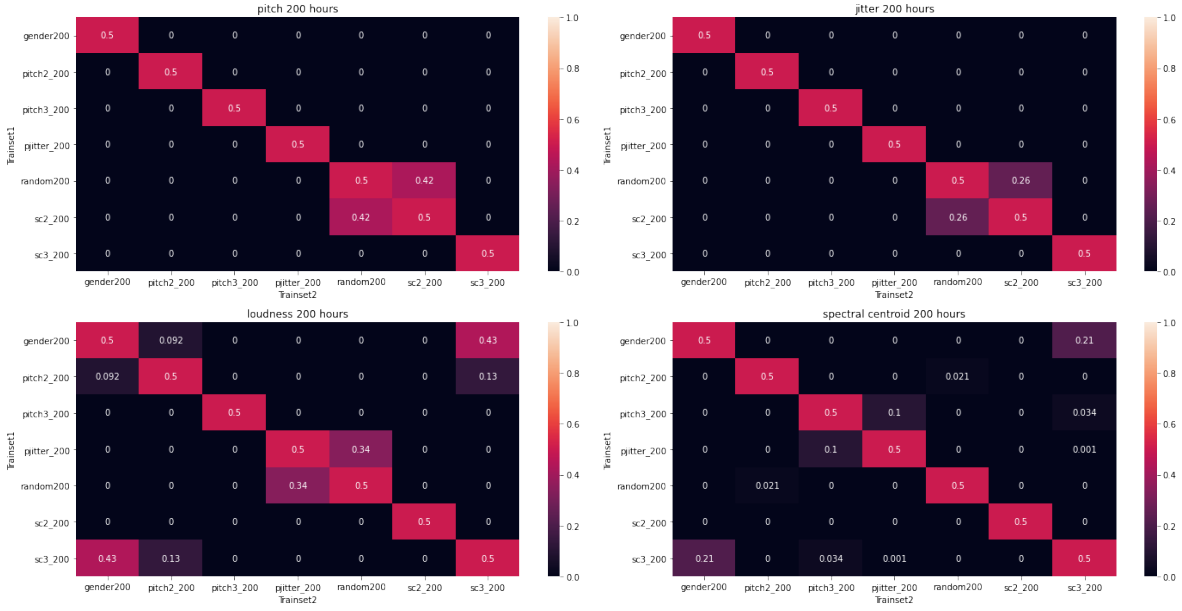

Figure 7.10: Mann-Whitney U p-values for each pair of 200 hours train sets and acoustic feature.

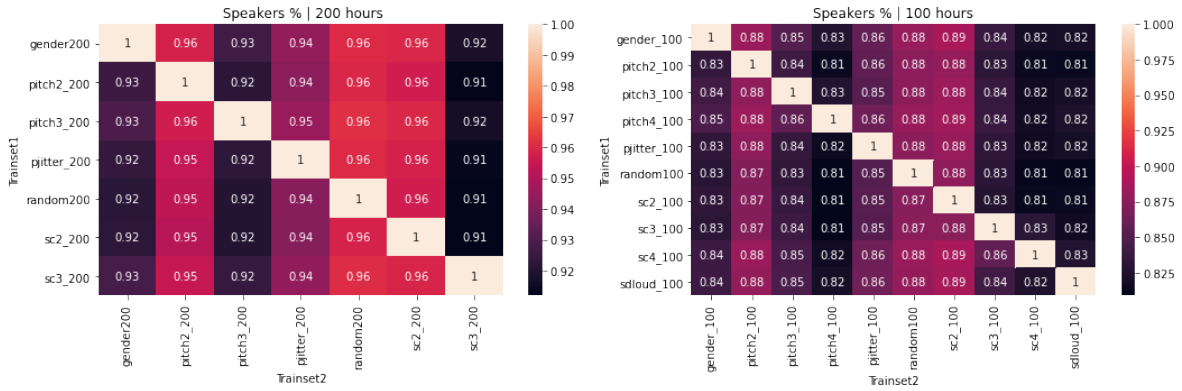

Figure 7.11: Percentage of speakers repeated for each pair of train sets.

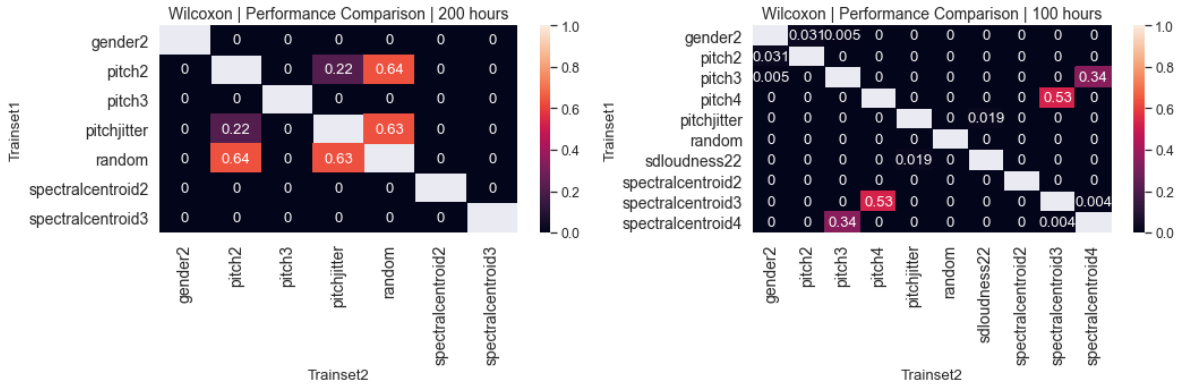

Figure 7.12: Wilcoxon performance results - 100 and 200 hours train sets.

Balancement Criterion	P-values	
	100H	200H
Gender	1.23E-62	1.50E-73
Pitch2	2.89E-71	4.45E-79
Pitch3	7.20E-77	1.17E-79
Pitch4	2.81E-65	NaN
Pitchjitter	1.06E-73	2.91E-83
Random	9.90E-68	1.53E-80
Sdloudness22	3.67E-80	NaN
Spectralcentroid2	2.59E-69	2.21E-78
Spectralcentroid3	4.42E-70	2.79E-80
Spectralcentroid4	5.40E-69	NaN

Table 7.2: Mann-Whitney U results, p-value results per balancement criterion.

Bibliography

[1] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2019.

[2] Nicol Turner Lee, Paul Resnick, and Genie Barton. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms, 5 2019. URL https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

[3] Smriti Parsheera. A gendered perspective on artificial intelligence. In *ITU Kaleidoscope 2018 – Machine Learning for a 5G Future*, pages 1–7, 11 2018. doi: 10.23919/ITU-WT.2018.8597618.

[4] A. Reis, D. Paulino, H. Paredes, I. Barroso, M. J. Monteiro, V. Rodrigues, and J. Barroso. Using intelligent personal assistants to assist the elderlies an evaluation of amazon alexa, google assistant, microsoft cortana, and apple siri. In *2018 2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW)*, pages 1–5, 2018. doi: 10.1109/TISHW.2018.8559503.

[5] Mahault Garnerin, Solange Rossato, and Laurent Besacier. Gender representation in french broadcast corpora and its impact on asr performance. page 3–9, 2019. doi: 10.1145/3347449.3357480.

[6] John Garofolo, Lori Lamel, William Fisher, Jonathan Fiscus, David Pallett, Nancy Dahlgren, and Victor Zue. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993. URL https://hdl.handle.net/11272.1/AB2/SWVENO.

[7] E. Holliman, J. Godfrey, and J. McDaniel. Switchboard: telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520, Los Alamitos, CA, USA, mar 1992. IEEE Computer Society. doi: 10.1109/ICASSP.1992.225858. URL https://doi.ieeecomputersociety.org/10.1109/ICASSP.1992.225858.

[8] Diogo Botelheiro, Alberto Abad, João Freitas, and Rui Correia. Nativeness assessment for crowdsourced speech collections. pages 21–25, 03 2021. doi: 10.21437/IberSPEECH.2021-5.

[9] Ravichander Vipperla. Automatic speech recognition for ageing voices. 11 2011.

[10] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the*

United States of America, 117(14):7684–7689, apr 2020. ISSN 10916490. doi: 10.1073/pnas.1915768117. URL https://www.pnas.org/content/117/14/7684https://www.pnas.org/content/117/14/7684.abstract.

[11] David Gerhard. Audio Signal Classification : History and Current Techniques Department of Computer Science University of Regina. (May), 2014.

[12] James P. Allen. Phonemes and Phones. *Ancient Egyptian Phonology*, 1:73–84, 2020. doi: 10.1017/9781108751827.008.

[13] Rita Singh. *Profiling Humans from their Voice*. Springer, 01 2019. ISBN 978-981-13-8402-8. doi: 10.1007/978-981-13-8403-5.

[14] M. Sambur. Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(2):176–182, 1975. ISSN 0096-3518. URL https://www.academia.edu/32689928/Selection{_}Of{_}Acoustic{_}Features{_}For{_}Speaker{_}Identification.

[15] Garima Sharma, Kartikeyan Umapathy, and Sridhar Krishnan. Trends in audio signal feature extraction methods. *Applied Acoustics*, 158:107020, 2020. ISSN 1872910X. doi: 10.1016/j.apacoust.2019.107020. URL https://doi.org/10.1016/j.apacoust.2019.107020.

[16] D. O'Shaughnessy. Linear predictive coding. *IEEE Potentials*, 7(1):29–32, 1988. doi: 10.1109/45.1890.

[17] Ameer A Badr and Alia K Abdul-Hassan. A Review on Voice-based Interface for Human-Robot Interaction. *Iraqi Journal for Electrical and Electronic Engineering*, 2020. doi: 10.37917/ijeee.16.2.10.

[18] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87 4:1738–52, 1990.

[19] KM Ravikumar, R Rajagopal, and HC Nagaraj. An approach for objective assessment of stuttered speech using mfcc. In *The international congress for global science and technology*, page 19, 2009.

[20] Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. Rasta-plp speech analysis technique. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 121–124, USA, 1992. IEEE Computer Society. ISBN 0780305329.

[21] Xavier Valero and Francesc Alias. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE Transactions on Multimedia*, 14(6):1684–1689, 2012. doi: 10.1109/TMM.2012.2199972.

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018. doi: 10.1109/ICASSP.2018.8461375.

[23] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011. doi: 10.1109/TASL.2010.2064307.

[24] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. 10 2020. doi: 10.21437/Interspeech.2020-2650.

[25] Cyril Pernet and Pascal Belin. The role of pitch and timbre in voice gender categorization. *Frontiers in Psychology*, 3:23, 2012. doi: 10.3389/fpsyg.2012.00023.

[26] Maxwell Hope and Evan Bradley. Vocal pitch and intonation characteristics of those who are gender non-binary. 2019. doi: 10.13140/RG.2.2.17233.68961.

[27] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Ng. Deepspeech: Scaling up end-to-end speech recognition. 12 2014.

[28] Miroslav Novak. Evolution of the asr decoder design. volume 6231, pages 10–17, 09 2010. ISBN 978-3-642-15759-2. doi: 10.1007/978-3-642-15760-8_3.

[29] Douglas B. Paul and Janet M. Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, page 357–362, USA, 1992. Association for Computational Linguistics. ISBN 1558602720. doi: 10.3115/1075527.1075614. URL https://doi.org/10.3115/1075527.1075614.

[30] Daren C. Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1), 2008. ISSN 13548565. doi: 10.1177/1354856507084420.

[31] Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane. *Big Data and Social Science: A Practical Guide to Methods and Tools*. Chapman amp; Hall/CRC, 2016. ISBN 1498751407.

[32] Sahil Verma and Julia Rubin. Fairness Definitions Explained. *IEEE/ACM International Workshop on Software Fairness*, 18, 2018. doi: 10.1145/3194770.3194776. URL https://doi.org/10.1145/3194770.3194776.

[33] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

[34] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning, 2018.

[35] Harini Suresh and John V. Guttag. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002, 2019. URL http://arxiv.org/abs/1901.10002.

[36] Raphael Lenain, Jack Weston, Abhishek Shivkumar, and Emil Fristed. Surfboard: Audio feature extraction for modern machine learning, 2020.

[37] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, M. Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *LREC*, 2020.

[38] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[39] Ekapol Chuangsuwanich. *Multilingual techniques for low resource automatic speech recognition*. PhD thesis, 01 2016.

[40] Winston Haynes. *Wilcoxon Rank Sum Test*, pages 2354–2355. Springer New York, New York, NY, 2013. ISBN 978-1-4419-9863-7. doi: 10.1007/978-1-4419-9863-7_1185. URL https://doi.org/10.1007/978-1-4419-9863-7_1185.

[41] Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. A review of ASR technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction, WOCCI '09*, 2009. ISBN 9781605586908. doi: 10.1145/1640377.1640384.