# Representation Learning for Voice Profiling

Daanish Ali Khan

CMU-CS-19-125
August 2019

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

**Thesis Committee:**
Bhiksha Raj, Advisor
Rita Singh, Advisor

*Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.*

**Keywords:** Voice profiling, representation learning, deep learning

# Abstract

Voice-profiling is the deduction of a speakers characteristics from their voice, a problem that has many applications in audio forensics, law enforcement, security and health-care. Speaker characteristics that can determined include the speakers gender, age, and ethnicity along with other physical and demographic characteristics.

Prior work on computational voice-profiling techniques modelled the production of voice as a physical system, and defined multiple voice signal features that encode speaker characteristics. Recent advances in artificial neural networks has resulted in an improvement in performance across voice profiling tasks, but such methods are often purely data-driven; the representation and relationships between voice and speaker characteristics are learned from a large dataset, not necessarily leveraging the knowledge-based voice features from prior work.

We identify the key challenges of modern voice profiling as being: 1) learning a representation that captures the complex relationship between voice and speaker parameters, 2) designing a representation that is resilient to real world noise, and 3) learning a representation that is generalizable across recording conditions and speaker characteristics.

In this work, we combine domain-specific signal-processing features with state of the art neural network techniques to learn a generalizable audio representation for voice-profiling. The learned representation is evaluated on multiple voice-profiling tasks including prediction of speaker gender, native language, and geographical origin. We experimentally show significant improvements in real world performance of voice profiling using our proposed speech representation.

# Acknowledgments

I thank my advisors Professor Bhiksha Raj and Professor Rita Singh for their unwavering guidance and support.

I also thank my friend and colleague Mahmoud Al Ismail for his help throughout the course of this project.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

In this thesis, we study the problem of computational voice profiling. We review and evaluate the performance of knowledge-driven techniques as well as modern machine learning based, data driven methods for voice profiling. Based on the current state of voice profiling, we identify advantages and shortcomings of each approach. We propose a novel voice-profiling method that combines knowledge and data driven techniques, and experimentally evaluate its real world predictive performance.

## 1.1 Voice Profiling and Speaker Parameters

Voice profiling is the task of computationally deducing personal characteristics and information about a human speaker from their voice. There are several types of personal characteristics, or *speaker parameters*, relevant to computational voice profiling. Physical parameters include information such as the speaker's gender, weight and height. Other parameters capture the speaker's physiological state, including information like the speaker's age and heart rate. Demographic information about the speaker includes parameters like ethnicity, geographical origin and native language. Psychological characteristics include the speaker's emotional state and sociological parameters. Medical parameters measure the speaker's general state of health, level of intoxication, presence of neurological disorders, etc.

Applied voice profiling has several applications in law enforcement, security, and healthcare. Advances in artificial intelligence and machine learning has resulted in a large increase in research interest in the field. Computational voice profiling is based on the fundamental assumption that if any parameter influences the production of voice, the relationship can be discovered. The production of voice, however, is a complex bio-mechanical process with high variability. A speaker producing the same sounds, under the same controlled conditions, can result in a signal that is different in spectral and temporal representations. This high degree of variability in the human vocalization process presents a set of challenges intrinsic to computational voice profiling; any profiling system will have to be resilient to the inherent variations in the signal.

Speaker characteristics can directly (or indirectly) cause variations in the produced signal.

These effects can be observed and studied to discern the relationship between speaker parameters and the resulting change in the produced voice signal. The methods of identifying these relationships can broadly be split into two categories: knowledge-driven and data-driven approaches.

## 1.2 Knowledge-Driven Profiling

In the knowledge driven approach to voice profiling, characteristics of the voice signal that are known to be statistically significant indicators of speaker parameters are identified. These signal characteristics are based on the mechanical characteristics of the voice production process. This process has been studied extensively, and our current understanding of it is significantly detailed. Knowledge-driven profiling techniques mathematically model the production of voice as a physical system. Features in the voice signal are designed to capture characteristics of specific parts of the voice production system. The relationship between the proposed features and the speaker parameters they capture are studied under controlled environments to establish statistical significance. The vast amount of prior studies on knowledge-driven signal properties has resulted in the identification of several signal features that can be used to determine various speaker parameters.

Knowledge-based features do not necessarily capture the entire relationship between the voice signal and speaker parameter due to the complex nature of their interactions. Furthermore, many knowledge based features are highly susceptible to noise; this could potentially result in the incorrect extraction of features from the signal. For real-world voice profiling applications, this is a significant challenge due to the difficulties associated with procuring noise-free speech samples.

## 1.3 Data-Driven Profiling

Data-driven profiling techniques employ machine learning and statistical pattern recognition techniques to automatically identify signal properties that are indicative of speaker parameters. This process typically involves a large amount of training data, with speaker parameters labelled. Advances in machine learning have produced an improvement in voice profiling performance, with the recent advent of deep learning and artificial neural networks at the current forefront of voice profiling techniques. Neural networks are function approximators that are capable of learning highly complex relationships between the input voice domain, and speaker parameters.

These data-driven approaches are still susceptible to the inherent variability in the human voice. Often, this can lead to incorrectly identifying features as being statistically indicative of speaker parameters when they are in fact not related. This can happen due to machine learning and statistical models overfitting on a particular dataset. A major challenge in data-driven voice profiling, is enforcing generalizability of a model. Neural networks learn to extract their own feature representations from the data; these features are near-impossible to disambiguate.

This leads to data-driven techniques falsely identifying relationships and features based on the datasets used. Often, these false relationships are caused by models relying on dataset-specific characteristics like the recording conditions or a biased and non-representative distribution of speaker parameters.

In this study, we analyze the relative strengths and weaknesses of traditional knowledge-driven methods and the more recent data-driven techniques, and identify key challenges in the field of computational voice profiling. Evaluation techniques are designed specifically to measure the performance of various models and techniques with respect to the identified challenges. We propose a novel representation for voice profiling tasks that combines the scientific background of the knowledge-based methods, and the learning capabilities of state-of-the-art data-driven techniques. We evaluate the performance of our proposed techniques, and present results that strongly indicate significant improvement over existing methods.

# Chapter 2

# Background

In this section, we provide a brief overview of the bio-mechanical process by which human voice is produced. We discuss the relationship between speaker parameters and their effect on the voice production mechanism. We review prior work that identifies patterns in the voice signal that captures these effects, and prior work that leverages large amounts of data to learn new patterns. Finally, we identify three key challenges in the field of voice profiling today.

## 2.1 Production of Voice

In order to understand how speaker parameter estimation from voice is even possible, we need to understand the process by which voice is produced. The mechanism that produces voice is composed of three subsystems in the vocal tract. The vocal tract (Figure 2.1) consists of the laryngeal cavity, the oral cavity, the nasal cavity and the pharynx. The production of voice can be split into three subsystems: the air pressure system, the vibratory system, and the resonating system.

### 2.1.1 Air Pressure System

The ability to produce voice begins with the expulsion of air from the lungs. This provides and regulates air pressure to the vocal tract that is required by the vibratory system to produce sound. The amount of air flow in the vocal tract is determined by a combination of factors related to the speaker's lungs, diaphram, chest muscles, ribs, and abdominal muscles.

### 2.1.2 Vibratory System

The vibratory sub-system is entirely contained within the larynx, or voice box. The larynx is an organ in the top of the neck, and houses the vocal folds (Figure 2.2). The vocal folds are folds of tissue, composed of twin infoldings of three distinct tissues. They are attached to muscles in the larynx that allows a speaker to change their shape and tension. When airflow from the air pressure system passes over the vocal folds, they oscillate by moving towards (and away) from each other. This oscillation causes modulations in the air pressure wave, producing audible

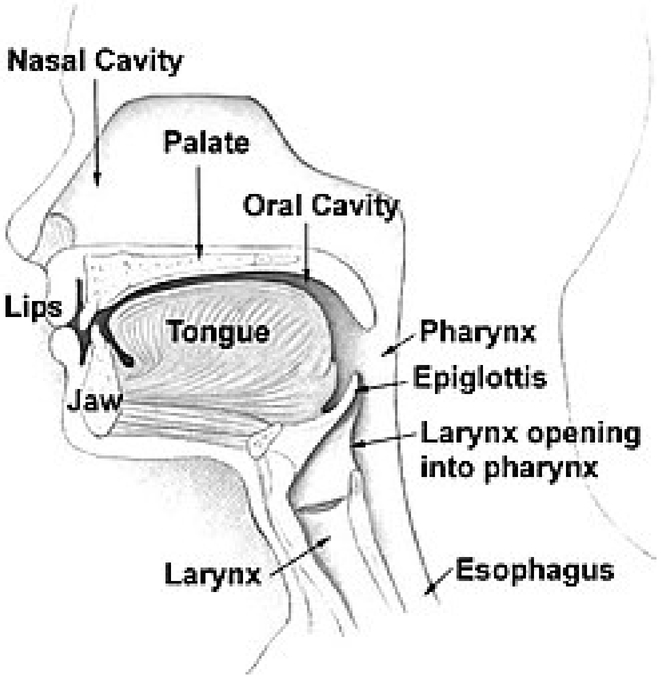

Figure 2.1: The Vocal Tract

sound. The tension and shape of the vocal folds determine the fundamental frequency at which they oscillate, determining the pitch of the produced sound. The muscles in the larynx control the fundamental frequency, allowing a speaker to produce sounds at different pitches. The human brain controls the voice production through specific nerve connections to the muscles in the larynx.

### 2.1.3 Resonating System

The sound that is produced by the Vibratory system is frequently perceived as being ”buzzy” in nature. The sound travels through the vocal tract to the resonating system, which consists of the pharynx, oral cavity and the nasal cavities. These act as resonators, as the sound produced in the larynx forms standing waves within each cavity. The specific resonances of each cavity is determined by their morphological structure, i.e their size and shape. The collective resonances transform the buzzy sound produced by the vocal folds into a speaker’s recognizable voice. *Articulators* in the mouth consist of the lips, tongue, jaw and teeth. These form the voice into specific voiced sounds, allowing speakers to produce specific phonemes. The voiced sound that is produced is determined by the configuration of the articulators, a series of specific configurations inside a speaker’s mouth is required to produce a spoken word. The size and shape of the articulators, as well as the configurations they can take, affect the speaker’s voice.

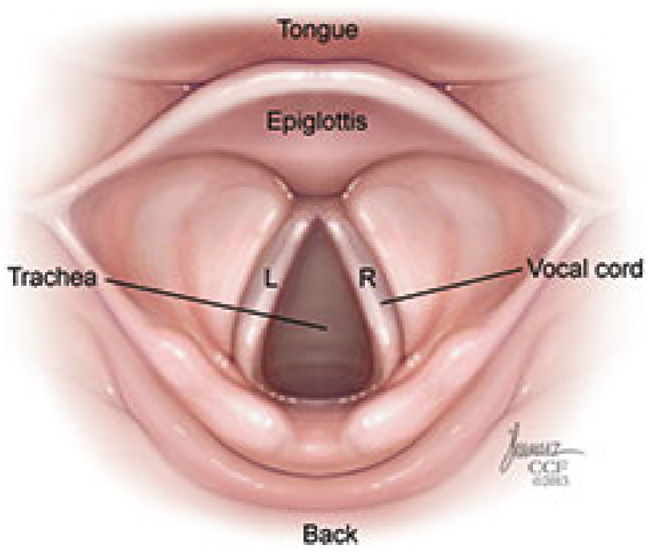

Figure 2.2: The larynx and the vocal folds

## 2.2 Basis of Voice Profiling

Speaker parameters affect the organs that are used in each subsystem of human speech production. These effects cause the produced sound wave to have different characteristics, and these changes can be observed in the produced voice signal. The fundamental basis of voice profiling is that speaker parameters produce specific observable patterns in the voice signal, and by analyzing these patterns one can determine the speaker parameters[15].

In the previous section we saw that there are many small sub-components in the voice production system. Each sub-component's shape, size and function affects the produced voice. Voice profiling studies the relationship between the speaker parameters discussed in Chapter 1, and how they affect these subcomponents and the voice signal. An example of one such relationship is the role of speaker gender in voice. The average length of the male vocal tract is approximately 17cm, while the average female vocal tract is only 14cm long. This results in changes in the behaviour of their resonating systems, specifically resonance at a higher frequency for females than males. The effect this difference has on voice, is the produced signal will have higher frequencies of resonance, or a higher pitch. This potentially allows us to estimate gender from the pitch.

## 2.3 Prior Work

The prior work on computational voice profiling can be broadly split into two categories: knowledge driven and data driven.

### 2.3.1 Knowledge-Driven Profiling

Knowledge-driven profiling techniques [1, 4, 6, 7, 8, 13] model the voice production system mathematically. They try to define mathematical characteristics of the voice signal that capture the behaviour of specific sub-components of the voice production mechanism. These mathematical characteristics, or signal features, are then statistically correlated with measurements of the sub-components and the speaker parameters they are believed to estimate. The correlation is established in carefully controlled environments, and are shown to be statistically significant. Since these methods capture the behaviour of human voice production, they are generalizable. For example, the mathematical formula to estimate the fundamental frequency of vocal fold oscillation is the same for all speakers, as all speakers produce voice using the same fundamental system. Small variations in the sub-systems from speaker to speaker will result in different estimated pitches, but we don't need to model each individual speaker's voice production separately to estimate their pitch.

As stated in Chapter 1, while the signal features that are used in knowledge-driven techniques can accurately estimate speaker parameters from clean speech samples, they are highly susceptible to noise. In reality, there are several external factors that influence the voice signal after it leaves the speaker, and these changes can adversely affect the estimated parameters. Figure 2.3.1 shows the spectrogram and estimated pitch (blue points) on a clean speech sample, and the same sample with synthetically added background babble noise. The estimated pitch is different, even though the frequency of vocal fold vibration was exactly the same in both signals. An estimator that relied on pitch to determine speaker parameters might have made an incorrect prediction because of the noisy pitch estimates.

### 2.3.2 Data-Driven Profiling

Recent advances in the field of Deep Neural Networks (DNNs) has led to significant performance improvement in voice-profiling tasks [2, 3, 9, 10, 11, 12, 14]. Neural networks are universal function approximators that have been shown to be highly effective at capturing complex and nuanced relationships in data. They have been used extensively as tools for audio feature extraction and classification tasks; sophisticated network architectures have outperformed all other methods and models on profiling tasks. This improvement in performance indicates that while the traditional spectral features are based on scientific observations of the bio-mechanics of voice production, they fail to capture the entire relationship between voice and speaker characteristics.

Neural networks are the best performing models for a variety of voice-profiling tasks including speaker identification, speaker verification, age estimation [11, 14], gender prediction [2, 3] and facial image reconstruction [16]. Such networks are trained on large labelled datasets containing several hundred hours of speech and speaker parameters.

The relationship between the voice signal and the speaker parameters is often learned in a purely data-driven fashion, not necessarily utilizing the signal features that prior work deter-

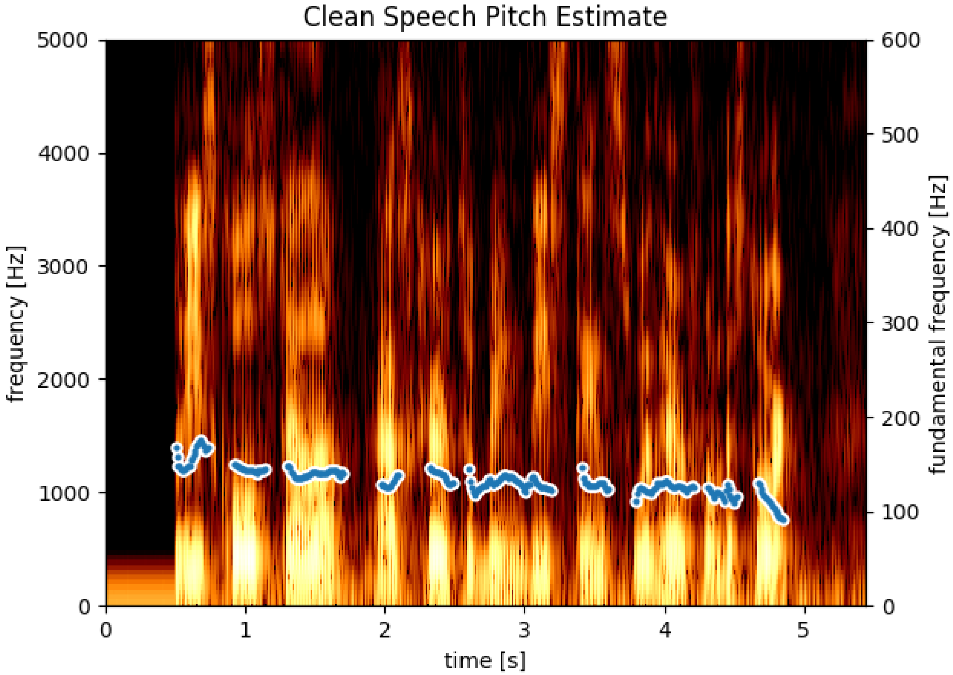

Figure 2.3: Clean speech sample

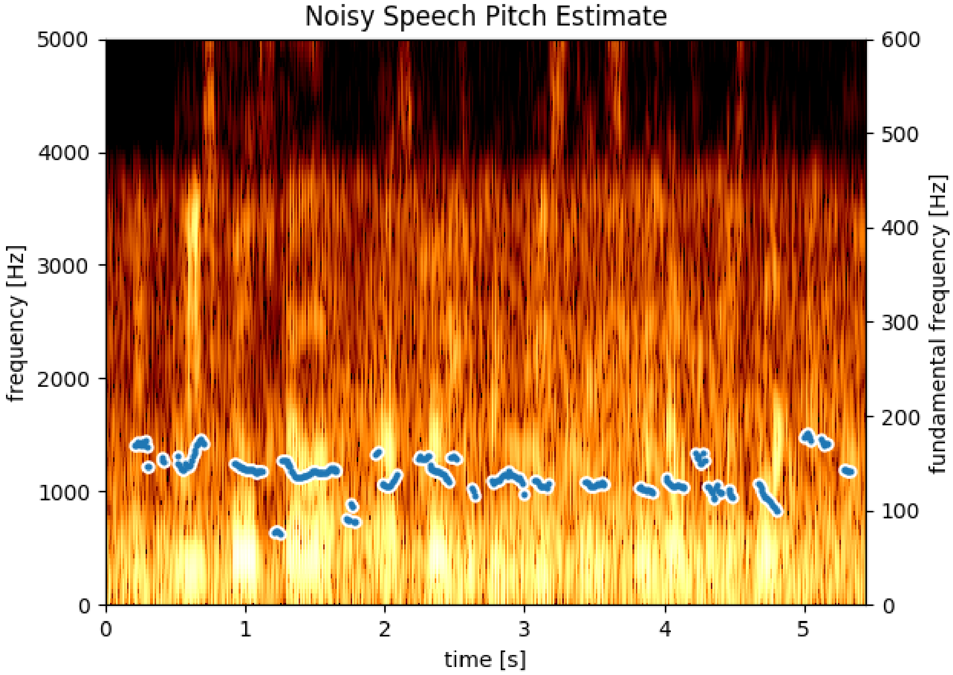

Figure 2.4: Noisy speech sample

mined to be statistically significant for voice-profiling. Furthermore, due to the complex nature of audio signals, neural networks tend to *overfit* on datasets they are trained on. This occurs when the model learns relationships between the input signal and speaker characteristics that are a peculiarity of the dataset, i.e the learned mapping might be true for one dataset, but not true in general. This leads to sufficiently high profiling accuracy on one dataset that does not scale to other data. One of the major challenges in deep learning for voice-profiling is designing generalizable models and representations such that the performance across datasets is maintained.

## 2.4 Key Challenges and Goals

From the limitations of knowledge-driven and data-driven methods, we can identify the three key challenges associated with the current state of voice profiling: predictive performance, noise resilience, and generalizability. In order to learn a representation of human voice that can accurately predict speaker parameters in real-world scenarios, we must address these three challenges.

### 2.4.1 Noise Resilience

In order for a model to be feasible for deployment, it will need to be able to perform accurately even in the presence of noise. We have seen that knowledge driven features are highly susceptible noise related issues, and estimations of signal features can be incorrect. Data driven methods are more robust to noise, but still suffer a decrease in accuracy.

### 2.4.2 Generalizability

The generalizability of a representation measures its ability to capture the true relationship between speaker parameters and voice signal. A model that learns false relationships will not be able to generalize to data that is outside the conditions of the false relationships learned. We have seen that knowledge-driven features are designed to capture the voice production system's parameters, and in the absence of noise, these features are highly generalizable. Data driven techniques have been shown to over-fit on specific recording conditions or speaker distribuitions of their training corpora, and do not generalize well without fine-tuning or extreme amounts of data.

### 2.4.3 Predictive Performance

The overall predictive performance of a representation can be defined as the amount of speaker parameter information that is captured within it. This can be measured by evaluating a representation on voice profiling tasks using real-world recordings. From prior work, we know that knowledge driven features have good predictive performance, but the predictive performance of data-driven models is significantly higher based on achieved accuracy on datasets. This is a combination of the noise resilience and the generalizability of a model. We can also measure the predictive performance by assessing the accuracy of a representation at estimating multiple speaker parameters.

# Chapter 3

# Methods and Contributions

In this chapter, we describe the datasets and tasks used to evaluate voice profiling representations and their performance on the key challenges identified in the previous chapter. We define a subset of the knowledge-driven signal features uses in the work, and provide baseline architectures to capture the features' raw predictive performance. We also provide a baseline data-driven architecture and evaluate its performance. We propose a novel fusion of data-driven and knowledge-driven techniques, and describe multiple iterations of models implemented. We evaluate all models on our benchmark tests, and report the overall performance breakdown.

## 3.1 Datasets and Benchmarks

One of the largest limiting factors for voice profiling is the availability of data. In order to train any predictive model that uses a data-driven technique, a large number of recordings is required. Furthermore, these recordings need to be provided with their speaker's labelled parameters. While there is an abundance of speech and voice data, speaker parameter labels are scarce. There are very few datasets that satisfy both the requirements of having a large number of recordings and multiple parameters labelled.

Throughout this study, we extensively use the Accent Archive Dataset[**?** ], a database established to uniformly exhibit a large set of speech accents from a variety of speaker backgrounds. The dataset contains over 2,100 recordings of speakers from 177 countries. The dataset contains speaker labels for *gender*, *native language*, and *geographical origin*. We use a subset of this dataset to train our models to predict these three speaker parameters from voice.

In order to measure the general performance of the trained models, we use the remaining data (hold out test set) from the Accent Archive Dataset to predict parameters and compare to ground truth. Three versions of the test set were used, stratifying the recordings based on each of the three labelled parameters.

To measure the resiliency of the models to noise, we created a new version of this test set, and augmented the data with noise consisting of a mixture of many voices, ambience from a

| Task | Baseline Accuracy(%) |
|---|---|
| Gender (AAD) | 50 |
| Gender (VoxCeleb) | 50 |
| Gender (Real-World) | 50 |
| Native Language (AAD) | 0.67 |
| Geographical Origin (AAD) | 0.83 |

Table 3.1: Random Baseline Performance

restaurant, and ambience from a street. Each recording was randomly augmented with one such noise type. The performance of the parameter prediction on this test set captures the resilience of a model to noise signatures that are representative of real world scenarios [5].

In order to measure the generalizability of models, we use the VoxCeleb dataset[**?** ], a collection of celebrity interviews with labelled speaker genders. It is important to note that the data used to train all models was entirely from the Accent Archive Dataset; as such the model would need to be highly generalizable to achieve high gender classification performance on this dataset. Standard practice in data-driven techniques involves fine-tuning the model on the new dataset, but this compromises our ability to measure generalization.

In addition to these two datasets, we were provided independent access to a dataset of over 600 real world recordings with gender labels. These speech samples were recorded under very noisy conditions, and capture the true conditions of real-world recordings. Throughout the remainder of this thesis, we refer to this dataset as Real-World. Similar to the VoxCeleb dataset, we only use these recordings to test and evaluate performance; no model was trained on this data. Due to the noisy nature of recordings, variety of speaker demographics, and lack of training data under those exact conditions, we believe that predictive performance on this dataset is highly representative of the true performance of a model.

As such, we used a benchmark suite of 8 distinct tasks. These tasks were designed to measure noise resiliency, representation generalization, and overall predictive performance. Table 3.1 shows the performance of the random baseline for each task. The random baseline for each task is to simply predict the most likely class, regardless of the input[1].

## 3.2 Signal Features for Voice Profiling

Extensive studies on voice-profiling techniques in the past established that the voice signal is significantly affected by minute changes in the mechanisms for voice production. Furthermore, empirical evidence of changes in the voice signal has been linked to several speaker parameters. Rigorous studies linking such physical parameters to signal properties provide a scientific basis

[1] Note that for the grographical origin and natice language tasks, the number of possible classes is much higher than that of gender. As a result the baseline accuracy for both is $< 1\%$

for voice profiling, supported by stastically valid findings.

Voice forensics studies have linked specific characteristics of voice and speech to the physical structure and configuration of the vocal-tract and the bio-mechanical process of voice production. Several studies have independently found that the physical attributes of a speaker directly impact the spectral characteristics of the voice signal.

The relationships between spectral properties and the speaker's physical parameters were studied using mathematical models that represent the production of speech as a physical system. The vast amount of prior work in this field has resulted in a large number of signal properties being identified as information-rich features for voice-profiling. It has been shown that many characteristics of the voice signal are affected by the physical parameter of speaker gender. Variations of the length of the vocal tract cause significant differences in the fundamental frequency of the speech signal.

Physiological parameters are those that relate to the biochemical function of the living body. Age is considered to be a physiological parameter, and has been shown to affect the speech signal's pitch, formants and spectral energy measures. Existing literature has shown that estimation of age based on spectral features alone is possible.

Demographic parameters like geographical origin has also been shown to affect the voice signal. This is due to childhood exposure to a language, which causes the formation of speaking habits that are often unique to the native language. This affects the speaker's coordination of vocal articulators, and directly influences what sounds and phonemes they are able to produce. These habits result in specific patterns of spectral characteristics that encode information about the speaker's geographical origin and their native language. Prior studies have shown that spectral features, such as the first and second formant spaces of vowels, are particularly strong indicators of such demographic parameters.

## 3.3 Signal Features

In this study, we implement signal feature extraction in order to benchmark their efficacy as voice-profiling features. In this section, we define the signal features used and analyze the performance of classifiers trained on such features.

### 3.3.1 Zero and Mean Crossing Rates

In signal processing theory, a *zero crossing* is said to occur when the signal value changes sign, while a *mean crossing* occurs when the signal value crosses the average value of the signal. The

Zero Crossing Rate ($ZCR$) of a signal $s$ over a segment of $T$ samples is defined as:

$$ZCR = \frac{1}{T-1} \sum_{n=0}^{T-1} \frac{|s_t s_{t+1}| - s_t s_{t+1}}{2|s_t s_{t+1}|} \tag{3.1}$$

Where $s_t$ is the $t^{th}$ sample in the signal. ZCR is easily affected by external noise, as such it is better suited to clean (non-noisy) speech recordings.

### 3.3.2 Signal Energy

The enrgy of a signal is defined by the area under the squared magnitude of the signal. Due to the discrete representation required for computational analysis of audio signals, the energy $E_s$ of the signal $s$ can be mathematically defined as:

$$E_s = \sum_{t=-\infty}^{\infty} |s_t|^2 \tag{3.2}$$

The energy contour of a signal is the sequence of energy values computed over fixed window periods. This feature is said to encode information about the spoken content of the signal, as well as speaker characteristics.

### 3.3.3 Spectral Centroid

The spectral centroid of a signal is an indication of where the center of mass of the spectrum is located. It is believed to be related to the perceived *brightness* of a sound. The spectral centroid is computed on the frequency domain representation of the signal, i.e. the signal is decomposed via the Fourier transform. The centroid is computed by taking a weighted average of the frequency components, using their respective amplitudes as the weighting function. It is believed to measure the timbre of the speech signal, encoding information about the speaker's characteristics.

### 3.3.4 Spectral Slope

The spectral slope of an speech signal is defined as the difference in the amplitude of specific harmonics in the speech spectrum, and has been widely believed to be indicative of glottal flow dynamics. It has been shown to be an information rich feature for voice-profiling. The spectral slope is computed by fitting a linear regression model on a segment of the Fourier magnitude spectrum, and using the gradient of the line of best fit.

### 3.3.5 Harmonics to Noise Ratio

The Harmonics to Noise Ratio (HNR) is a feature that compares the relative energies in the harmonics in the speech spectrum, to the energy in the noise (non-harmonic) portions. The region of speech harmonics is generally believed to lie within the frequency range of $70Hz$ to $1500Hz$, while the noise region is within $1500Hz$ to $4500Hz$. The HNR of a speech signal is said to capture aspects of voice quality, and can encode significant information relating to the speaker's characteristics. Naturally, the presence of additive noise adversely affects the HNR, making it better suited to clean speech recordings.

### 3.3.6 Pitch

The pitch of an audio signal is a key feature that correlates several speaker characteristics. The fundamental pitch frequency $F0$ is defined as the fundamental frequency of phonation. This can be measured directly fom the voice signal using the *autocorrelation method.*

The autocorrelation method segments the speech signal into analysis frames of length 20-30ms. For each analysis frame, the autocorrelation function is defined as:

$$R_{ss}[\tau] = \frac{1}{N-\tau} \sum_{t=0}^{N-1-\tau} s_t s_{t+\tau} \tag{3.3}$$

The pitch period (inverse of the fundamental pitch frequency) is defined as:

$$\tau_{\text{pitch}} = \arg \max_{\tau \in [\tau_{min}, \tau_{max}} R_{ss}[\tau] \tag{3.4}$$

## 3.4 Voice Profiling with Spectral Features

In order to evaluate the significance of speech signal features for the purpose of voice profiling, we performed a series of experiments to predict speaker characteristics from a vector of mean spectral features. Random Forest classifiers were used to perform the classification tasks. Random Forests are an ensemble machine learning classification technique that learns a multitude of decision trees during training time, and predicts a class based on a maximum voting scheme across decision trees. In our experiments, $100$ decision trees were trained for each ensemble classifier.

The voice features were evaluated on a set of three tasks. The first task was to predict the gender of the speaker from their voice features. This is considered to be a relatively simple binary classification problem, particularly when dealing with clean speech recordings, as prior research has shown that exceptionally high classification accuracy can be achieved. Studies have found that the gender of the speaker has significant impact on almost all aspects of the speech signal. This is due to the biological differences in the voice production mechanisms between males and females. Based on this, one may expect generally high classification accuracy for gender.

The second task is to predict the speaker's native language. Note that all recordings in our dataset contain exclusively English speech. The classifier will learn to map voice signal features to one of 214 native languages. The Accent Archive Dataset contains several languages from similar regions in the world; similar languages will result in similar phonation habits learned by native speakers in their childhoods. As a result, correctly classifying a speaker's native language is considered a challenging task.

The third task is to predict the speaker's geographic origins. Similar to the native language, the geographical origins of a speaker will likely influence the coordinaztion of their articulators and the phonemes they are capable of producing. The classifier will predict one of 177 geographical origins for each speaker.

| Profiling Task | Accuracy (%) |
|---|---|
| Gender (AAD) | 94.6 |
| Gender with noise (AAD) | 73.1 |
| Gender (VoxCeleb) | 86.8 |
| Gender (Real-World) | 59.8 |
| First Language | 20.5 |
| First Language (noisy) | 17.9 |
| Geographical Origin | 18.0 |
| Geographical Origin (noisy) | 11.7 |

Table 3.2: Classification accuracy using average signal features

## 3.5 Deep Learning and Spectral Features

Recent advances in the field of Deep Neural Networks (DNNs) has led to significant performance improvement in voice-profiling tasks. Neural networks are universal function approximators that have been shown to be highly effective at capturing complex and nuanced relationships in data. They have been used extensively as tools for audio feature extraction and classification tasks; sophisticated network architectures have outperformed all other methods and models on profiling tasks. This improvement in performance indicates that while the traditional spectral features are based on scientific observations of the bio-mechanics of voice production, they fail to capture the entire relationship between voice and speaker characteristics.

Neural networks are the best performing models for a variety of voice-profiling tasks including speaker identification, speaker verification, age estimation, gender prediction and facial image reconstruction. Such networks are trained on large labelled datasets containing several hundred hours of speech and speaker parameters.

The relationship between the voice signal and the speaker parameters is often learned in a purely data-driven fashion, not necessarily utilizing the signal features that prior work determined to be statistically significant for voice-profiling. Furthermore, due to the complex nature of audio signals, neural networks tend to *overfit* on datasets they are trained on. This occurs when the model learns relationships between the input signal and speaker characteristics that are

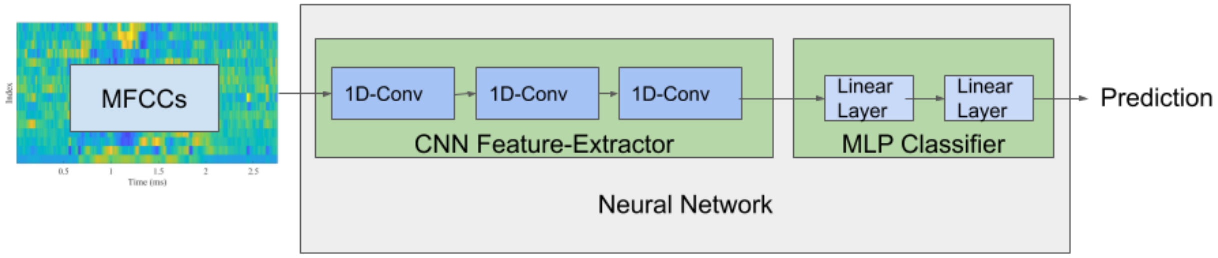

Figure 3.1: Baseline Network Architecture

a peculiarity of the dataset, i.e the learned mapping might be true for one dataset, but not true in general. This leads to sufficiently high profiling accuracy on one dataset that does not scale to other data. One of the major challenges in deep learning for voice-profiling is designing generalizable models and representations such that the performance across datasets is maintained.

## 3.6 Baseline Model

Typically, neural network models for voice analysis are not trained on the raw audio signal as input; instead, a frequency domain representation is used. For the purposes of comparison, we define a baseline neural network model for voice profiling as operating on the Mel Frequency Cepstrum Coefficients (MFCCs). MFCCs are a representation of the short-term power spectrum of sound. Unlike the regular frequency domain representation, MFCCs are based on a nonlinear mel scale of frequency (a scale of pitches that are perceived to be equal in distance from one another).

Our baseline model (3.1 for voice-profiling is a Convolutional Neural Network (CNN), a neural network architecture that is designed to extract shift invariant features from the input space. CNNs have been shown to perform exceedingly well on a number of audio analysis tasks, and is widely used in deep learning methods for voice profiling.

## 3.7 Neural Network Trained on Spectral Features

A naive method of incorporating signal features directly into a neural network architecture would be to pre-compute all the features in Section 3.1 and use them as input to a neural network classifier. If the baseline model was not utilizing all of these features in its internal intermediate representation, we would expect to see an improvement in performance across tasks.

The network architecture used in for this model (3.2) is identical to the baseline model, with the exception of the input dimensionality including multiple signal processing features in addition to the MFCC input of the baseline.

| Profiling Task | Accuracy (%) |
|---|---|
| Gender (AAD) | 97.0 |
| Gender with noise (AAD) | 79.4 |
| Gender (VoxCeleb) | 71.5 |
| Gender (Real-World) | 80.5 |
| Native Language | 32.1 |
| Native Language (noisy) | 27.8 |
| Geographical Origin | 38.5 |
| Geographical Origin (noisy) | 30.2 |

Table 3.3: Classification accuracy using baseline neural network

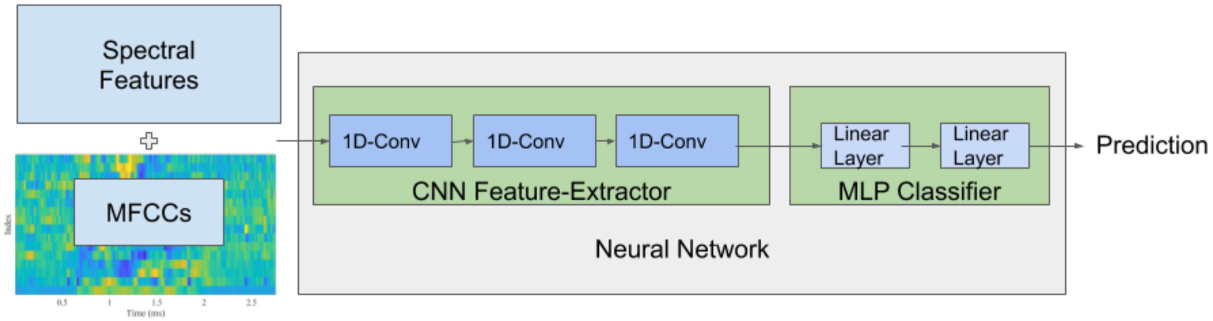

Figure 3.2: Spectral Feature Network Architecture

| Profiling Task | Accuracy (%) |
|---|---|
| Gender (AAD) | 98.5 |
| Gender with noise (AAD) | 77.0 |
| Gender (VoxCeleb) | 80.1 |
| Gender (Real-World) | 79.1 |
| Native Language | 37.5 |
| Native Language (noisy) | 31.5 |
| Geographical Origin | 42.5 |
| Geographical Origin (noisy) | 33.3 |

Table 3.4: Classification accuracy using signal features and baseline neural network

## 3.8 Spectral Feature Reconstruction Network

In this section, we present a novel neural network framework for voice profiling. Our goal when designing this network, is to incorporate the signal features into the internal representation of our neural network, without incurring the higher computational costs associated with the increased input dimensionality. In order to do this, we can no longer perform expensive pre-processing tasks during inference. Thus, the input to our model will be the same as that of the baseline model, i.e just the MFCC representation.

A common technique employed in neural network representation learning is to incorporate some form of reconstruction from the intermediate representation. A naive approach would be to reconstruct the input signal from a fixed length feature vector, and penalize the network based on the reconstruction error. This form of reconstruction, however, will result in the network encoding the content of the speech (what is being said) in its representation. For the purposes of voice profiling, the speech content is not of high significance; ideally the network will only encode the quality of the speech (how it is being said). This is due to the fact that the speech quality is what holds the most information regarding the speaker's characteristics.

From the vast amount of prior work on voice quality features derived from the voice signal, and the performance of these features on profiling tasks, it is clear that the features described in Section 3.1 do encode speaker characteristics, and therefore voice quality. Based on this knowledge, we propose a modified reconstruction loss, that instead of reconstructing the entire speech signal, reconstructs the signal features of the original input from the intermediary representation. Note that the signal features are not passed as input to the network, forcing the model to learn a representation of the audio signal that directly encodes this information.

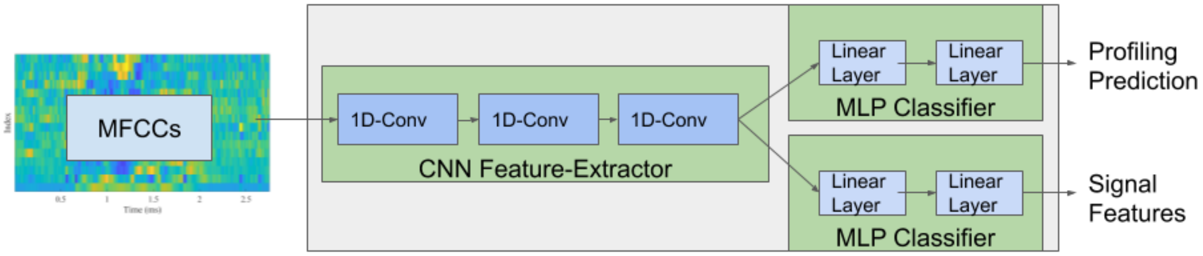

Figure 3.3: Spectral Reconstruction Network Architecture

In order to improve the performance and robustness of the spectral feature reconstruction network, resiliency to noise must be learned. As shown in Figure 2.3.1, the signal features are highly susceptible to noisy speech.

A commonly used technique in deep learning based speech processing, is noise augmentation. The training data is modified to contain synthetically added noise. This allows the network to learn a representation that can separate the noise from the speech signal, improving robustness to noise significantly. In order to address the sensitivity of the signal features to noise, we extract the features from the clean signal and train the network to reconstruct them from the noisy speech signal. The architecture of this model is the same spectral feature reconstruction network, the only difference being the MFCCs are computed from the artificially noisy speech signal. This allows the network to learn the true voice signal features from noisy data, making the representation resilient to noise while still encoding the desired voice quality features.

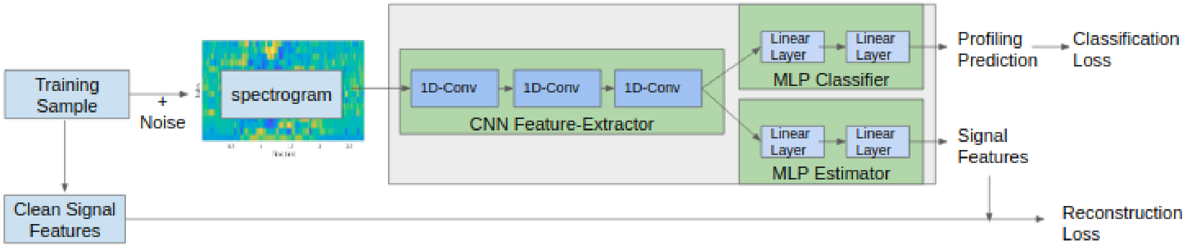

Figure 3.4: Augmented Signal Reconstruction Training System

| Profiling Task | Accuracy (%) |
|---|---|
| Gender (AAD) | 98.4 |
| Gender with noise (AAD) | 95.0 |
| Gender (VoxCeleb) | 92.5 |
| Gender (Real-World) | 93.8 |
| Native Language | 44.0 |
| Native Language (noisy) | 43.6 |
| Geographical Origin | 47.2 |
| Geographical Origin (noisy) | 45.0 |

Table 3.5: Classification accuracy using spectral reconstruction network

## 3.9 Multi-Task Learning for Voice Profiling

Prior work on voice profiling has shown that speaker parameters' effects on the voice signal are rarely independent from one another. This means that one speaker parameter will inflcuence the patterns caused by other speaker parameters. An example of this is the effect of age on voice. As a person ages, their voice changes. The way the voice changes, however, depends on the gender of the speaker. Thus, to accurately estimate the age of a speaker, the representation must also encode the gender.

Our methods thus far do not directly incorporate multiple parameters into our learned representation. Furthermore, by using one model for each speaker parameter, the computational costs at inference time for multiple parameters becomes unscalable for any real-world deployment of a system. A technique that can be used to solve both these problems is Multi-Task Learning. This formulation trains one representation to perform multiple related yet different parameter prediction tasks, forcing the representation learned to encode multiple speaker parameters (Figure 3.5). This also reduces the computational costs associated with inference as only one feature representation is needed.

We modify the architecture of our signal reconstruction network to allow for the multi-task learning (Figure 3.5, and train the model to predict gender, geographical origin, and native language jointly.

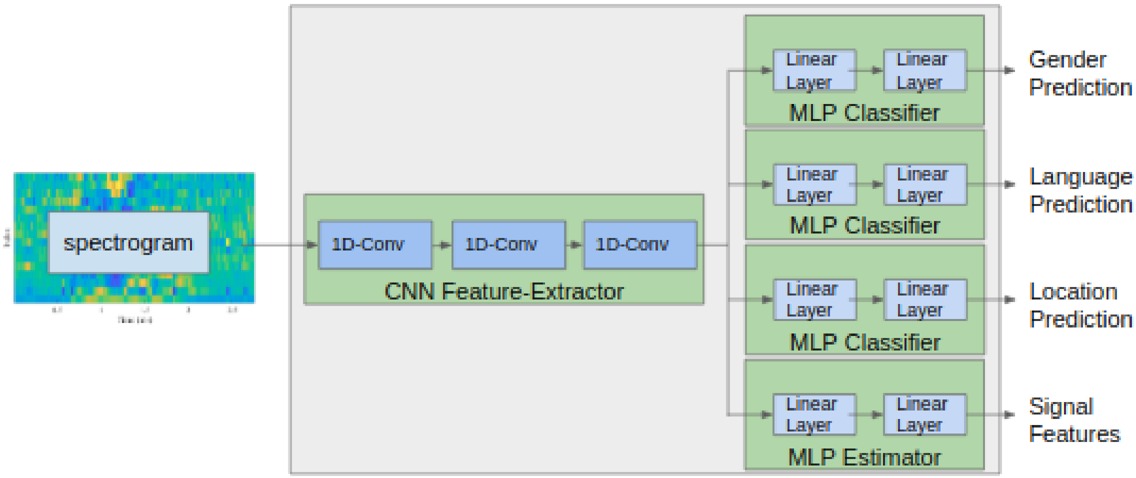

Figure 3.5: Architecture of Multi-Task Learning Network

| Profiling Task | Accuracy (%) |
|---|---|
| Gender (AAD) | 98.7 |
| Gender with noise (AAD) | 95.5 |
| Gender (VoxCeleb) | 94.0 |
| Gender (Real-World) | 97.8 |
| Native Language | 45.4 |
| Native Language (noisy) | 44.0 |
| Geographical Origin | 49.7 |
| Geographical Origin (noisy) | 47.3 |

Table 3.6: Classification accuracy using multitask network

# Chapter 4

# Discussion and Conclusions

In this chapter we discuss the benchmark results of the various architectures and models described in Chapter 3, and evaluate overall performance of our proposed architectures. We outline the key conclusions we can draw from the results, and identify limitations of our work and future research directions.

## 4.1 Analysis of Benchmark Results

### 4.1.1 Random Forest

The knowledge driven features used to train the random forest were hypothesized to be more generalizable, and less resilient to noise. The results in Figure 4.1 support this claim. We see that the performance on the VoxCeleb dataset is relatively high when compared to the purely data-driven results in Figure 4.2. We also note that the classification accuracy drops significantly on the noisy datasets. This also indicates that the features used are highly susceptible to noise in the input signal.

### 4.1.2 Neural Network Baseline

The results of baseline model for the neural network trained on the frequency domain representation of the input signal can be seen in Figure 4.2. Based on the purely data-driven methods' limitations, we believe that the network is unable to generalize easily across datasets. This can be seen in the accuracy on the VoxCeleb dataset, the performance drops considerably from the Accent Archive Dataset (AAD) gender test set. This model appears to be relatively more resilient to noise in the input, however a decrease in performance is still observed. Performance on the real world data does not match the performance on the AAD test set, also indicating a lack of generalizability.

### 4.1.3 Fusion Network Baseline

The results of the network trained on the combination of the signal features and the frequency representation can be seen in Figure 4.3. Due to the incorporation of the knowledge driven

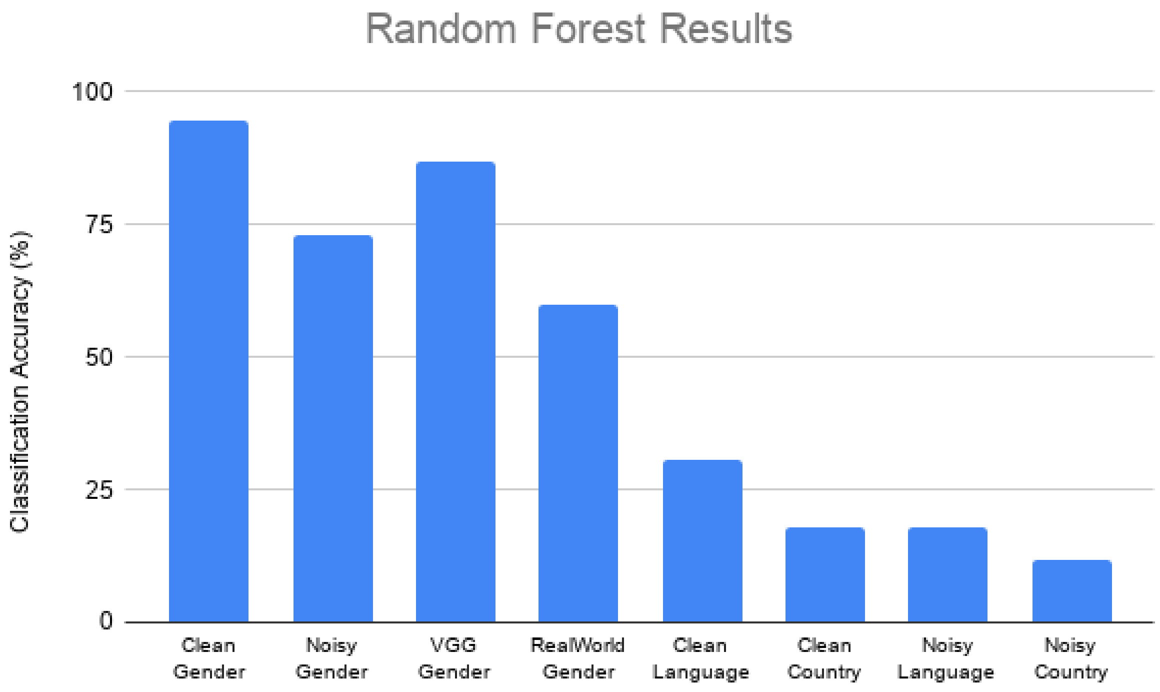

Figure 4.1: Benchmark results for knowledge-driven random forest baseline

features, we expect the network to learn a more generalizable representation, while at the same time being less resilient to noise. The results on the VoxCeleb dataset support our generalizability hypothesis, and the decreased performance on the noise augmented test sets support our noise susceptibility hypothesis. Overall predictive performance of this architecture has improved over the knowledge-driven and data-driven baselines. This indicates that a fusion of the two methods allows the representation to learn more complex relationships between the input and speaker parameters, while still being generalizable.

### 4.1.4 Signal Reconstruction Network

The results of the augmented reconstruction network can be seen in Figure 4.4. In this model, we expected the noise resiliency to improve significantly, along with an improvement in generalizability. We expect the network to retain its performance on the overall predictive performance tasks characterized by the accuracy on the clean AAD gender, geographical origin, and native language.

The results support our claims regarding noise resiliency and generalization of the representation. The performance on the noisy dataset drops by a much smaller margin, indicating strong resilience to noise. This is due to the addition of our noise augmented training scheme. We also see a significant improvement on the VoxCeleb dataset, supporting the claim that the learned representation is capable of generalizing across datasets.

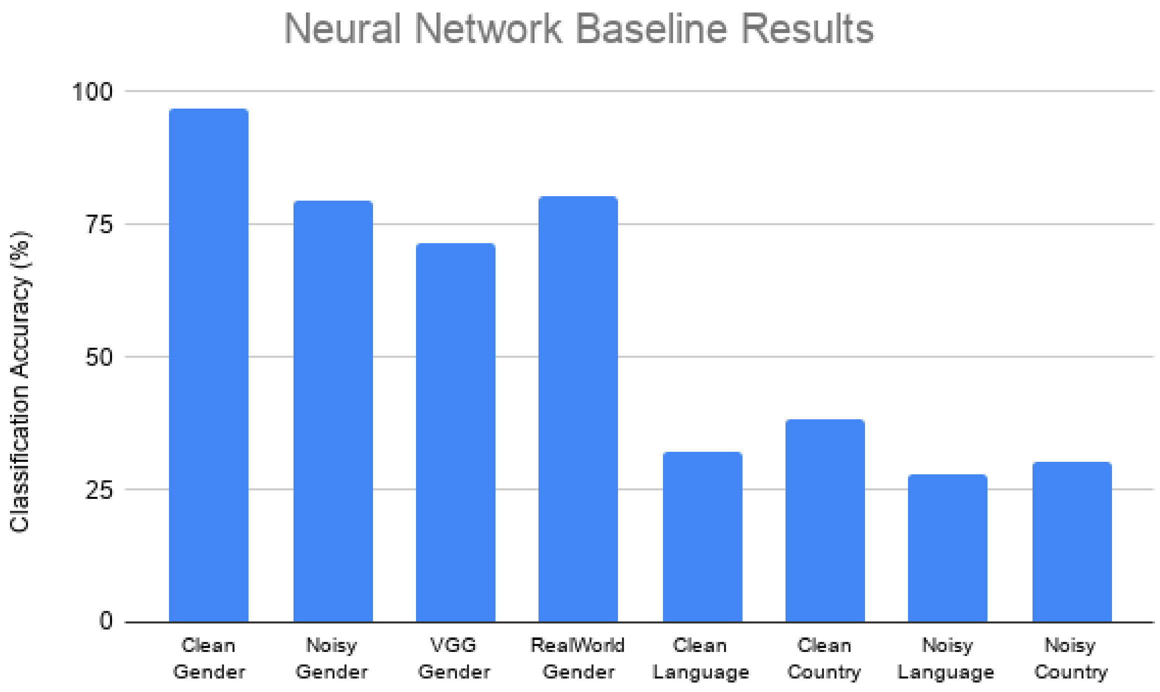

Figure 4.2: Benchmark results for neural network baseline

There was an improvement in performance for native language and geographical origin prediction, contrary to our expectations. This result is likely due to the fact that the model learned a more information-rich representation of the input signal while trying to reconstruct the signal features. This representation captures an even more complex relationship than previous models; the network extracts voice signal features from the noisy spectrogram input, and uses this representation to predict speaker parameters. Our hypothesis is that the network has learned a general representation of the clean voice signal features that may include more information than just the knowledge-driven features of the clean signal.

### 4.1.5 Multi-Task Network

The results of the multi-task signal reconstruction network are reported in 4.5. By incorporating the multi-task training, we expect the overall predictive performance to improve. The results support this, with an improvement in all benchmark tasks. The improvement in performance is due to the internal representation encoding multiple speaker parameters, thus allowing the prediction of one parameter to leverage information about other parameters that are encoded in the embedding space. We also observed an improvement in performance on the noisy recordings, while we expected the performance to remain the same. This improvement is due to the representation capturing overall better predictive performance, thus even in the presence of noise it is capable of making better predictions.

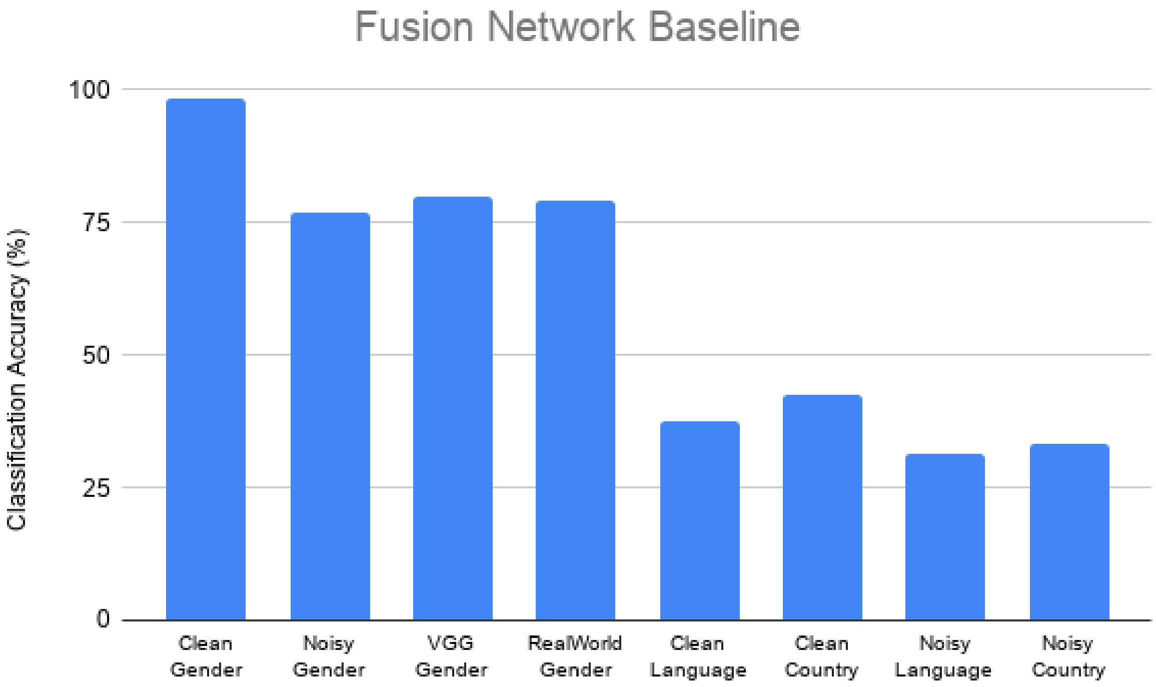

Figure 4.3: Benchmark results for neural network fusion baseline

## 4.2 Real-World Performance

Figure 4.6 compares the speaker gender prediction performance on the real-world data. We see that the knowledge-driven method achieves an accuracy only marginally better than the random baseline for the task. The poor performance of this model can be attributed to the highly noisy nature of the recordings, and the noise susceptibility of derived signal features. The features that were passed to the random forest classifier during inference were incorrectly estimated due to this noise.

The data-driven baseline achieves a higher accuracy on this data, largely due to the fact that the model is trained on the frequency representation of the signal directly. As a result, the presense of noise will not affect the performance to the same degree as the knowledge-driven approach.[1]

The naive knowledge and data driven fusion network achieves a lower accuracy on the real world data. This can be entirely attributed to the noise susceptibility of the signal features that the network is trained on; incorrect signal feature estimates are passed to the neural network, resulting in incorrect predictions.

[1] The network is not entirely resilient to noise, we have seen in Section 4.1.2 that the presence of noise does result in a decrease in performance.

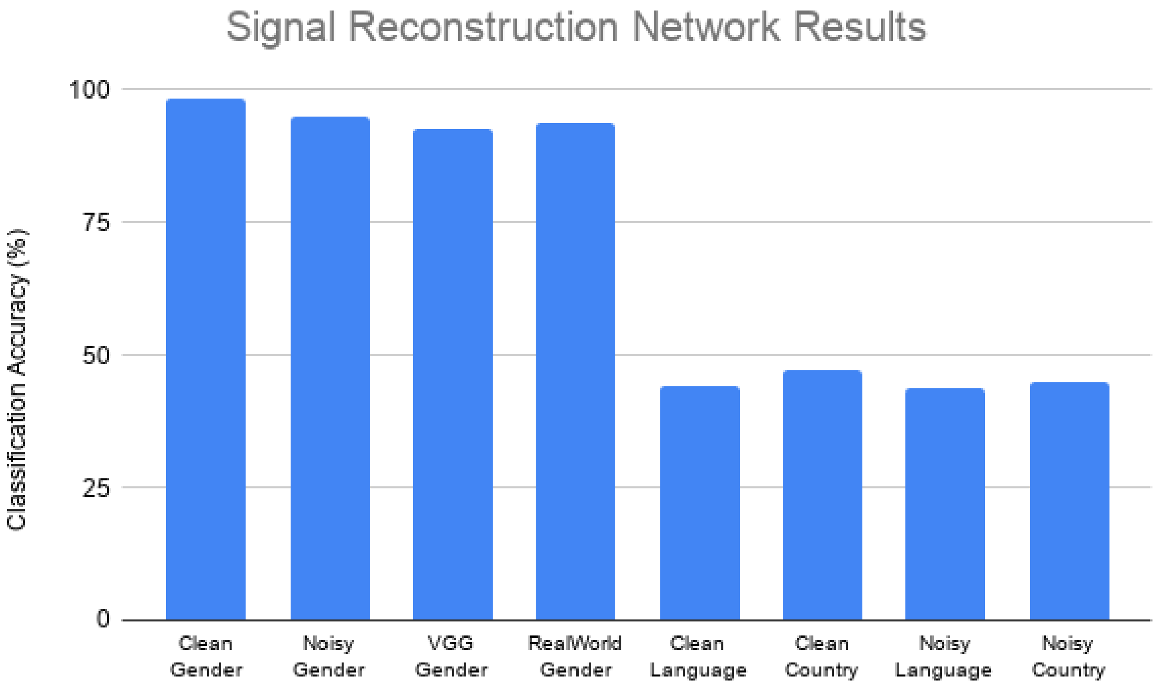

Figure 4.4: Benchmark results for signal reconstruction network

We see that our proposed signal reconstruction network achieves a much higher accuracy, nearing the predictive performance on the hold out test set (clean gender test set from AAD). This strongly indicates that the representation learned by this model is highly generalizable. The presence of real noise does not appear to decrease classification performance significantly, indicating that our noise augmentation during training successfully allowed our model to distinguish the speech components from the noise in the signal.

The multi-task signal reconstruction network achieves the best performance on the real world dataset, improving the accuracy from the single-task signal reconstruction network. This improvement in performance is attributed to the fact that the representation learned by the model, encodes multiple other speaker parameters. This indicates that the incorporation of native language and geographical origin information improves gender classification.

## 4.3 Future Work

One of the main limitations in this study was the availability of labelled data. Our methods did not necessarily address the issue of data-driven techniques related to the large data requirement. Future work on the topic should include larger labelled datasets for the task of voice profiling [2].

[2] At the time of this study, no such voice-profiling datasets exist.

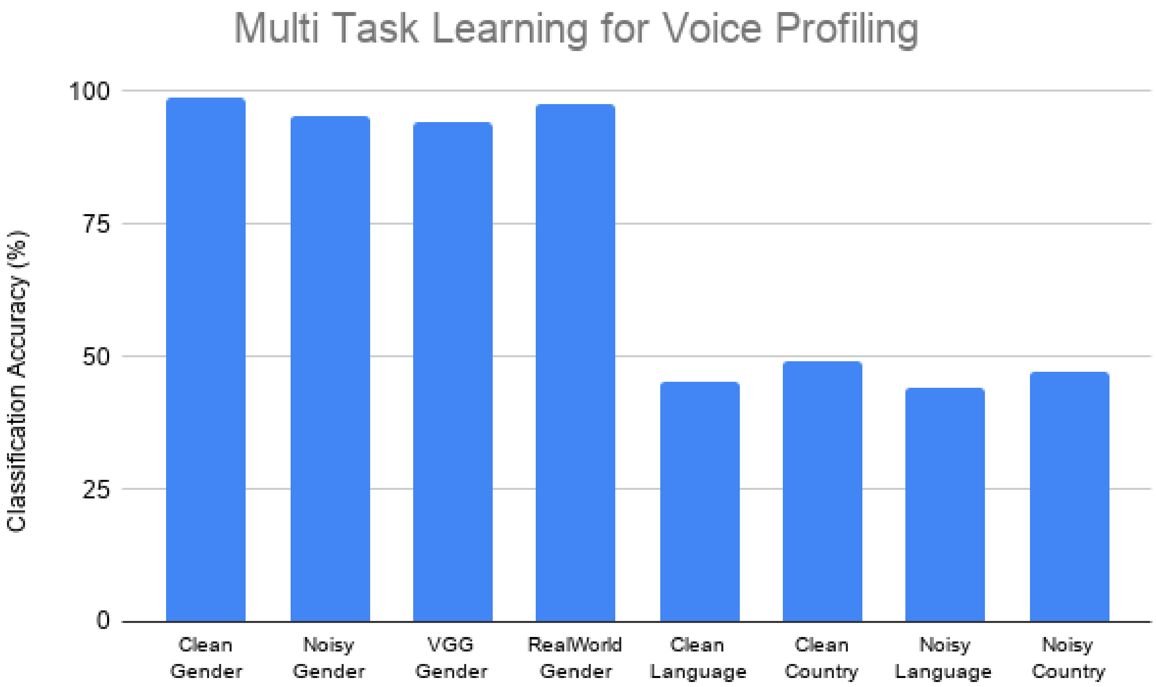

Figure 4.5: Benchmark results for multi-task network

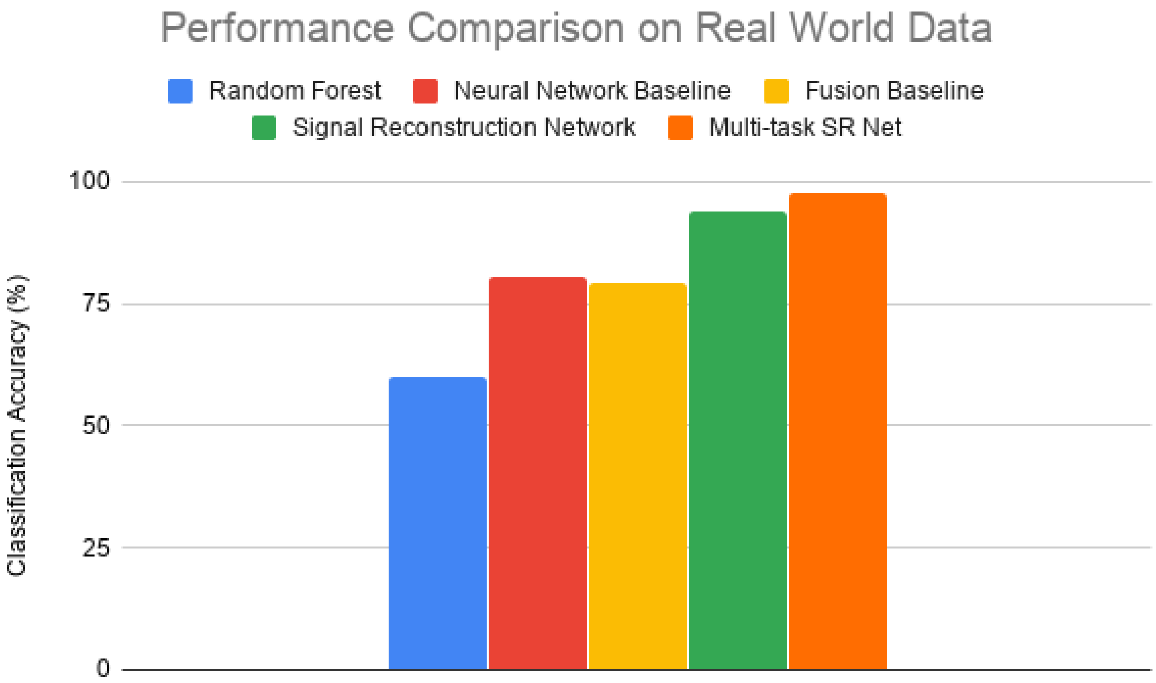

Figure 4.6: Real-World Gender classification comparison

In our research we utilized only a small subset of knowledge-driven signal and spectral features for reconstruction. A plethora of prior literature has identified a multitude of relevant features that could potentially improve performance significantly. Furthermore, an analysis of which signal features are most useful for each task can provide more domain-specific knowledge. The architecture proposed in this work will, in theory, scale to other signal features as well as voice quality parameters (parameters used to describe the perceived nature of the voice).

While we achieved an improvement in overall performance by using multi-task learning, we only used three application specific tasks (gender, geographical origin, and native language) due to data limitations. The incorporation of more voice profiling tasks can further improve this performance.

## 4.4 Conclusions

By combining knowledge-driven features based on the production of voice, with sophisticated data-driven techniques like convolutional neural networks and multi-task representation learning, we were able to significantly improve voice profiling performance. Based on prior work, we identified three key challenges in the current state of voice profiling: noise resilience, generalizability, and predictive performance, and designed mechanisms to address each one. We proposed a method of benchmarking the performance of a model relative to these challenges, and experimentally showed that our solutions addresses the shortcomings of the current state of the art voice profiling methods.

In this work we proposed a novel neural network architecture to incorporate signal features into an internal representation without using them as input. We augment our input with noise during training allowing the network to learn to separate the noise from the clean speech signal.

We have shown that even though state of the art data-driven techniques report very high accuracy on hold-out test data, on their own they are not generalizable enough to deploy reliably in real world conditions. Using a test set consisting of data recorded under real world conditions, we experimentally show that our proposed model is objectively better suited for deployment in a real voice profiling system.

We have shown that purely data-driven methods can be significantly improved by incorporating domain-specific knowledge into the learned representation, and presented a novel mechanism for doing so, improving performance across all relevant metrics.

# Bibliography

[1] Tarika Bhuta, Linda Patrick, and James D Garnett. Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of voice*, 18(3):299–304, 2004. 2.3.1

[2] Mucahit Buyukyilmaz and Ali Osman Cibikdiken. Voice gender recognition using deep learning. In *2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016)*. Atlantis Press, 2016. 2.3.2

[3] Shih-Hau Fang, Yu Tsao, Min-Jing Hsiao, Ji-Ying Chen, Ying-Hui Lai, Feng-Chuan Lin, and Chi-Te Wang. Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice*, 2018. 2.3.2

[4] Hanspeter Herzel, David Berry, Ingo Titze, and Ina Steinecke. Nonlinear dynamics of the voice: signal analysis and biomechanical modeling. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 5(1):30–34, 1995. 2.3.1

[5] Hans-Günter Hirsch and David Pearce. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000. 3.1

[6] Dennis H Klatt and Laura C Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, 87(2):820–857, 1990. 2.3.1

[7] Jody Kreiman, Bruce R Gerratt, Kristin Precoda, and Gerald S Berke. Individual differences in voice quality perception. *Journal of Speech, Language, and Hearing Research*, 35(3): 512–520, 1992. 2.3.1

[8] Jody Kreiman, Bruce R Gerratt, Gail B Kempster, Andrew Erman, and Gerald S Berke. Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research*, 36(1):21–40, 1993. 2.3.1

[9] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the iEEE conference on computer vision and pattern recognition workshops*, pages 34–42, 2015. 2.3.2

[10] Nobuaki Minematsu, Mariko Sekiguchi, and Keikichi Hirose. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–137. IEEE, 2002. 2.3.2

[11] Amir Hossein Poorjam, Mohamad Hasan Bahari, et al. Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 7–12. IEEE, 2014. 2.3.2

[12] Amir Hossein Poorjam, Mohamad Hasan Bahari, Vasileios Vasilakakis, et al. Height estimation from speech signals using i-vectors and least-squares support vector regression. In *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*, pages 1–5. IEEE, 2015. 2.3.2

[13] Klaus R Scherer. Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology*, 8(4):467–487, 1978. 2.3.1

[14] MH Sedaaghi. A comparative study of gender and age classification in speech signals. *Iranian Journal of Electrical and Electronic Engineering*, 5(1):1–12, 2009. 2.3.2

[15] R. Singh. *Profiling Humans from Their Voice*. Springer Nature Singapore Pte Limited, 2019. ISBN 9789811384035. URL `https://books.google.com/books?id=9_edDwAAQBAJ`. 2.2

[16] Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, and Rita Singh. Disjoint mapping network for cross-modal matching of voices and faces. *arXiv preprint arXiv:1807.04836*, 2018. 2.3.2