AUTOMATIC ESTIMATION OF PERSONAL CHARACTERISTICS USING SPEECH DATA

Thesis

Submitted in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

by

KALLURI SHAREEF BABU

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING,
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA,
SURATHKAL, MANGALORE - 575025

June 2021

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

DECLARATION

I hereby ***declare*** that the Research Thesis entitled **AUTOMATIC ESTIMATION OF PERSONAL CHARACTERISTICS USING SPEECH DATA** which is being submitted to the ***National Institute of Technology Karnataka, Surathkal*** in partial fulfillment of the requirement for the award of the Degree of ***Doctor of Philosophy*** in **Department of Electronics and Communication Engineering** is a ***bonafide report of the research work carried out by me***. The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

Kalluri Shareef Babu,
Reg. No.: 138014 EC13F07
Department of Electronics and
Communication Engineering.

Place: NITK-Surathkal.
Date: 6-7-2021

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

CERTIFICATE

This is to certify that the Research Thesis entitled **AUTOMATIC ESTIMATION OF PERSONAL CHARACTERISTICS USING SPEECH DATA** submitted by **KALLURI SHAREEF BABU** (Register Number: 138014 EC13F07) as the record of the research work carried out by him, is accepted as the *Research Thesis submission* in partial fulfillment of the requirements for the award of degree of **Doctor of Philosophy**.

Deepu

Dr. Deepu Vijayasenan
Research Guide
Dept. of E & C Engg.
NITK Surathkal - 575025

Chairman - DRPC
Department of Electronics and
Communication Engineering
(Signature with Date and Seal)

प्राध्यापक एवं विभागाध्यक्ष / PROF & HEAD
ई एवं सी विभाग / E & C Department
एन आई टी के, सुरतकल/N I T K, Surathkal
मंगलूर / MANGALORE - 575 025

Acknowledgements

I want to express my sincere gratitude to my research advisor Dr. Deepu Vijayasenan for guiding me throughout the research work. The valuable suggestions and guidance provided by him kept me on the right track. He is the one who has all answers for all my endless queries. He has a very great ability to spot my errors in both codings quickly as well as technically and corrects me at every point of time I struck. He was motivated at every fall time of mine, both technically and personally. I am indebted to him for his support, guidance, and encouragement during my research.

I express my gratitude to Prof. Muralidhar Kulkarni, Head of the department, Electronics and Communication Engineering during my enrollment for the Ph.D. program, Prof. M. S. Bhat, Prof. U. Shripathi Acharya, Prof. T Laxminidhi, and Prof.Ashwini Chathuvedi, Heads of the Department of E&C Engineering during my research work for their support, help, and encouragement.

I am grateful to my RPAC members, Prof.Sumam David, Dept. of E&C Engg. and Dr.P.Jidesh, Assistant professor, Dept. of MACS, for giving vital comments and suggestions throughout the research, which helped in improving the quality of research. I would like to extend my thanks to Prof.Sumam David for being my teacher during course work and clearing all my academic doubts and personal doubts as well. I would like to sincerely thank Dr.Sriram Ganapathy, Assistant Professor, Dept. of EE, IISc Bangalore, who guided and helped me at the hard times by providing the vital inputs during this Ph.D. journey.

I would like to show my gratitude to Dr.Raghavendra Bobbi, Dr.Arulalan, Dr.Shyamlal, Prof.Ramesh Kini for being such friendly and motivating me professionally as well as personally in this journey. I would like to thank Mr.Sanjeev Poojari, Mrs.Pushpalatha, Ms.Amitha, Ms.Prabha, Mr.Guru Tilak, Mr.Subramanya Karanth, Mr. Ratheesh, for helping me in solving my nontechnical doubts. I want to thank all the faculty members and staff of the E&C department, NITK Surathkal, for their assistance.

I would like to thank all the student volunteers of NITK Surathkal, IISc Bangalore, SVEC Tirupathi, KSRIET Tiruchungode, College of Engineering Thalassery for helping me during the process of data collection; without their help, it would have been very difficult for me to complete my data collection.

I would like to express my gratitude to all my friends and colleagues at NITK to encourage me in the good and the bad times to make a memorable and beautiful stay in NITK. Thanks to Praveen Ram Teke, G Ramu, Raghavendra MANS, Chethan, Vanjari Venkata Ramana, Jayaram, Ragesh, Ranjan, Shajahan, Mahesh, Prasad, Goutham, Gnaneshwarnadh Sathapathi, Shrutakirti (too many to list here but you know who you are!) for technical discussions and support during my research work.

Special thanks to Lakshmi– my trek partner, who showed me a trekking taste. Later I continued my journey into the woods whenever I get an opportunity. I would like to thank Yajunath and S Ramu for being with me at a phone call distance for all my good and bad times. I would like to extend my thanks to my childhood friends Arif, Khadar and Khaleel for their encouragement during my studies. I also need to mention a few other friends in this journey, Madhu K, for the support, motivation, and patience in listening to me in my bad times, Prashant from LEAP Lab, IISc Bangalore, for his consistent help during my data collection. I want mention my mentors during my M.Tech internship at National Atmospheric Research Laboratory (NARL),Dr. TVC Sarma, Scientist-SG, Mr. Jayant Chouragade, Scientist-SE, NARL, Tirupati, who ignited a spark to take the steps towards Ph.D.

I am deeply indebted to all my teachers throughout life who have guided, encouraged, and inspired me to grow in both technical and personal aspects.

I express my gratitude to the Ministry of Human Resource Development (MHRD), and Science and Engineering Research Board (SERB) Government of India, for providing financial assistance during the research. I would also like to thank TEQIP-II for supporting me financially during my international travel for the MLSP-2016 conference and SERB for covering my registration charges for ICASSP-2019.

I want to thank my parents K Babji and S Thofa, brother K Shakil Babu, sister K Sharmila, my wife Ayesha, my grandmother, for their continuous support, love, and encouragement. I am grateful to all my family members, their unwavering love and support throughout the research work. Thanks to my lovely nephew Wasif Shahansha (Lucky), and my niece Jassiah Juhi for bringing extra zeal and motivation to my work.

Finally, I would like to thank god for giving me good health, strength, and bliss during my research work.

Kalluri Shareef Babu

Dedicated to

To My Beloved Family

ABSTRACT

Many paralinguistic speech applications demand the extraction of information about the speaker's characteristics from as little speech data as possible. In this work, we explore the estimation of the speaker's multiple physical parameters from the short duration of speech in monolingual (English) and multilingual settings. This has applications in forensics as well as $e-$commerce. We explore different feature streams derived from the speech spectrum at different resolutions. Short-term log-mel spectrogram, formant features, and harmonic features are extracted for age and body build estimation (height, weight, shoulder size, and waist size) of the speaker. The statistics of these features accumulated over the speech utterance are used to learn a support vector regression model for speaker age and body build estimation. The experiments performed on the TIMIT dataset show that each of the individual features can achieve results that outperform the default predictor (prediction of the mean of test samples by blindly predicting the mean of training data without looking at the features) in height and age estimation. Furthermore, the estimation errors from these different feature streams are complementary, allowing the combination of estimates from these feature streams to improve the results further. The combined system from short audio snippets achieves a performance of 5.2 cm, and 4.8 cm in Mean Absolute Error (MAE) for male and female, respectively, for height estimation. Similarly, in age estimation, the MAE is 5.2 years and 5.6 years for male and female speakers.

We extend the same physical parameter estimation system to other body build parameters like shoulder width, waist size, weight, and height. We created two datasets for the speaker profiling task in a multilingual and multi-accent setting. Speech data is collected along with speaker parameter details (like height, age, shoulder size, waist size, and weight). A pilot dataset Audio Forensic Dataset (AFDS) with 207 speakers across 12 different native Indian languages has around 8 hours of native languages speech and around 9 hours of English speech data. Later, a bigger dataset NITK-IISc Multilingual Multi-accent Speaker Profiling (NISP) dataset has collected, and it has 345 speakers across five Indian languages as well as English. NISP dataset has around 25 hours of native languages speech data and 32 hours of English speech data. The system can estimate all the physical parameters and showed better improvement than the default predictor in the multilingual and multi-accent setting.

The duration analysis shows that the state-of-the-art results can be achieved using short utterances(around $1-2$ seconds) of speech data. To the best of our knowledge, this is the first attempt to use a common set of features for estimating the different physical traits of a speaker from short utterances.

An integrated end-to-end deep neural network architecture is proposed for joint prediction of all the physical parameters. A novel initialization scheme for deep neural architecture is introduced, which avoids a large training dataset requirement. On the TIMIT dataset, the system achieves an RMSE error of 6.85 and 6.29 cm for male and female height prediction. In the case of age estimation, the RMSE errors are 7.60 and 8.63 years for male and female, respectively. Analysis of shorter durations of speech reveals that the network only degrades around 3% at most with only 1 second of the speech input. Also, the performance saturates around 3seconds in predicting the height and age of a speaker using the TIMIT dataset. In the multilingual setting using collected datasets, the predicted error metrics are less than the default predictor except for female age prediction in both AFDS and NISP datasets. In male speakers, the system performance is less than the default predictor in height estimation of the NISP dataset.

Keywords : Speaker Profiling, Multilingual data, AFDS, NISP, Short duration, Physical Parameters, Audio Forensics.

Contents

List of Figures

List of Tables

ABBREVIATIONS

AFDS	Audio Forensics Dataset
ANN	Artificial Neural Networks
AR	Auto Regressive
DNN	Deep Neural Network
F_0	Fundamental Frequency
Fstats	First Order Statistics
GMM	Gaussian Mixture Model
LSSVR	Least Square Support Vector Regression
MAE	Mean Absolute Error
MFCC	Mel Frequency Cepstral Coefficients
NISP	NITK-IISc Multilingual Multi-accent Speaker Profiling
RMSE	Root Mean Square Error
SGR	Sub-Glottal Resonance
SVM	Support Vector Machines
SVR	Support Vector Regression
TMP	Target Mean Predictor
UBM	Universal Background Model
VTL	Vocal Tract Length

Chapter 1

Introduction

Human speech data not only contains information about the textual message being conveyed but also the characteristics of the speaker. While the former is typically used in Automatic Speech Recognition (ASR), the latter information is effectively used in speaker identification and speaker verification (Schuller *et al.* (2013)). The pictorial idea of speaker profiling using speech data is shown in Figure 1.1. Advances in digital speech processing now support application and deployment of a variety of speech technologies for human/machine communication. The human speech data contain information about the following questions.

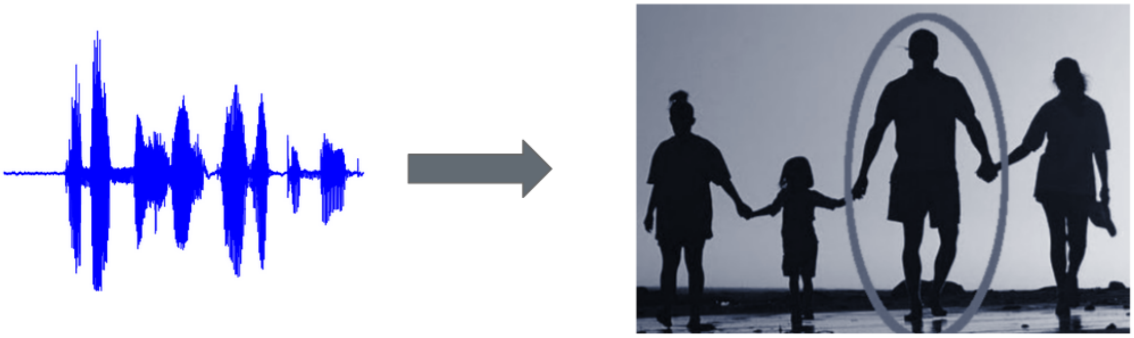

Figure 1.1: Pictorial idea in identifying the speaker attributes from the speech data

- **Who** – Speech data will help to identify ‘who’ the speaker is. Speech helps to determine the gender (male/female) of a speaker and to determine the age group of the person like a child, teenager, adult, senior citizen.

- **How** – Speech data will help in understanding the emotions of the speaker like anger, happiness, sadness, anxiety,etc., along with physical stature, weight, etc. of the speaker.

- **Where** – Speech data will help to identify the speaker, (region from where the speaker belongs to) using the accent information.

- **What** – The speech data has linguistic information, which can be transcribed into readable text, which is helpful in all automatic speech recognition applications.

- **Which** – The speech data has the linguistic content, which conveys the language that is spoken.

The extraction of speaker characteristics (parameters) from the speech data could further aid in speaker identification systems as well as in the speaker clustering and diarization systems. Speaker profiling involves predicting speaker meta information such as age, accent and parameters of body build like height, weight, shoulder size and waist size. Speaker profiling is a challenging application area (Tanner and Tanner (2004)). The main challenge in estimating any such information is the separation of linguistic content and speaker traits.

There are many potential applications in identifying the physical parameters from the speech data in developing the engineering systems for biometric applications (Nolan (2005), Singh *et al.* (2016*a*), Poorjam *et al.* (2015)) as well as commercial applications (Schuller *et al.* (2013)), forensics (Tanner and Tanner (2004)) etc.

The $e-$commerce applications like targeted advertisements on internet, caller-agent pairing in call-centers, video games etc would need more details on age and gender for attracting the specific group of people. The information about the users language/accent, age and gender can be used to offer appropriate products and services for the $e-$commerce applications. The knowledge about users characteristics can help in personalizing video games. For instance the choice of music is significantly different for a teenage boy from that of an adult or may be the case of a boy and a girl (Poorjam *et al.* (2015), Schuller *et al.* (2013), Singh *et al.* (2016*a*)).

Estimating the physical traits could supplement the voice forensic analysis in case of forensic scenarios besides providing knowledge to improve the speaker identification systems. Usually the forensic experts search for a list of suspects involved in these type of criminal activities. In hoax and threatening calls, speaker clues can be extracted from voice recordings. This is a manual and tedious time consuming task for the forensic experts to trace out the accused from the entire list of suspects. In such a scenarios, soft biometrics such as age, gender, accent, physical parameters etc., help to narrow down the number of suspects (Nolan (2005), Tanner and Tanner (2004), Singh *et al.* (2016*a*)).

There is no control over the amount of available speech data from the target speaker in such cases. Therefore, such systems are required to provide accurate predictions using a minimum amount of speech data. For example, DARPA RATS program targeted development of speaker and language recognition technology with as short as 3 seconds of speech (Walker and Strassel (2012)). Thus, development of speaker profiling methods in short duration audio is important. Most of the available resources for physical parameter estimation are based on mono language (mostly English). Often, in all the commercial, surveillance, forensic applications the available speech may not be in English. Thus, there is a need to develop a multilingual physical parameter estimation system from available short durations of speech data.

Motivated by the importance and need to estimate the speaker's physical parameters from short speech duration in a multilingual setting, for the commercial, surveillance, and forensic applications, we framed our thesis's objective as follows.

1.1 Objectives of the Thesis

- To investigate physical characteristics of a person from multilingual multi-accent speech data.

- Develop a system for reliable estimation of speaker characteristics from short duration of speech.

1.2 Contributions

The main focus of the thesis is to develop a common platform to estimate the physical parameters (Age, height, shoulder size, waist size, and weight) from short duration speech data. The main contributions of this thesis are summarized as follows:

- Exploration of different kinds of features that uncover the speech signal's underlying spectral structure to estimate the physical parameters. The statistics of Mel cepstral features, formants, and harmonics are extracted at the utterance level. This common set of features do not depend on phone level transcriptions. These features are able to achieve the state of the art results in height and age estimation tasks using the TIMIT dataset. The system is also able to retain similar performance even for short utterances(2s).

- Creation of two different multilingual and multi accent datasets – Audio Forensics Dataset (AFDS–207 speakers), and NITK-IISc Multilingual Multi-accent Speaker Profiling (NISP–345 speakers) from the native Indian speakers for estimating the physical parameters (like age, height, shoulder size, waist size, and weight) of a speaker.

- Extension of the physical parameter estimation to shoulder size, weight, waist size, along with height and age estimation of a speaker in a multilingual context. The multilingual system has minimal degradations as compared to the monolingual case.

- Proposal of an integrated end-to-end deep neural network to estimate the physical parameter with a limited amount of training data. The DNN predicts, a speaker's height and age jointly with short length utterances (1–3s) on the TIMIT dataset. The same end-to-end DNN architecture is used to predict jointly the height, age, weight, shoulder size, and waist size of a speaker in a multilingual setting with short duration utterances using AFDS and NISP datasets.

- A block diagram illustrating the challenges in speaker profiling and proposed methodology to address the challenges and the contributions made to the thesis are shown in Figure 1.2.

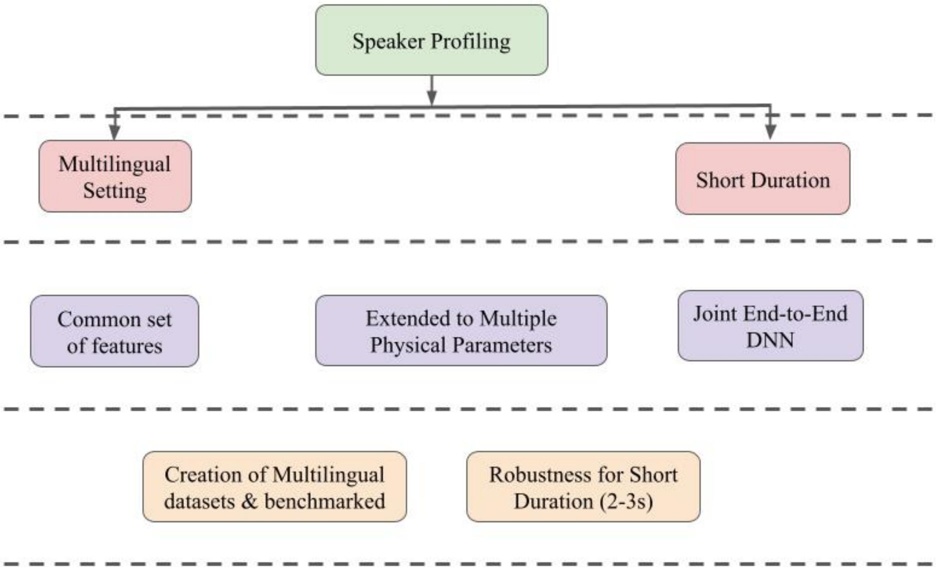
Speaker Profiling
Multilingual Setting
Short Duration
Common set of features
Extended to Multiple Physical Parameters
Joint End-to-End DNN
Creation of Multilingual datasets & benchmarked
Robustness for Short Duration (2-3s)

Figure 1.2: Diagram illustrating various contribution made in the thesis.

1.3 Outline

The rest of the thesis is organized as follows.

Chapter 2: Presents about prior work and background literature in physical parameter estimation using the speech data. This chapter is begins with the Physiological cues in speech to estimate the physical parameters. The chapter then reviews the speaker profiling literature. The chapter concludes with current challenges and motivations for this thesis.

Chapter 3: Details the exploration of different kinds of features (MFCC, formants and harmonics) that uncover the underlying spectral structure of the speech signal at multiple levels. We look for features that does not depend on the phoneme level transcriptions. Extensive experiments were carried out on the standard TIMIT dataset to predict the height and age of a speaker.

Chapter 4: Discusses about the two different multilingual and multi-accent speech datasets created for multiple physical parameter estimation. The detailed setup of data collection and the potential applications of the datasets are presented. The same approach discussed in chapter 3 is extended for estimating the multiple physical parameters estimation (height, shoulder size, waist size, weight and age) in a multilingual setting.

Chapter 5: Introduces an integrated end-to-end DNN architecture for joint estimation of multiple physical parameters using short duration speech data (1-3s). A novel scheme of initialization which eliminates the requirement of large amounts of training data is also discussed. This chapter details the experiments performed on the TIMIT dataset as well as collected multilingual and multi-accent datasets.

Chapter 6: Provides a general summary of the presented research work.

Chapter 2

Literature Review

Speaker profiling involves prediction of speaker meta information such as age, accent and parameters of body build like height, weight, shoulder size from the speech data.

The motivation for height estimation range from biological understanding of the anatomy and its relationship to the speech properties to development of potential engineering systems for biometric applications (Nolan (2005), Singh *et al.* (2016*a*), Poorjam *et al.* (2015)). While the current performance may not be applicable directly for developing robust solutions, the potential to augment speech based features as additional information has shown to improve other biometric methodologies based on finger printing (Jain *et al.* (2004)).

Researchers have focused on identifying a speaker's age group (children, youth, adult, and senior) from speech data rather than estimating the exact age. Most of the commercial applications (like targeted advertisements, caller-agent pairing in call-centers, etc) and forensic applications have focused in estimating the age of a speaker. Estimating the physical parameters could help to narrow down on suspects of hoax/threat calls, in forensic applications. (Nolan (2005), Singh *et al.* (2016*a*), Schuller *et al.* (2013)). Estimating the physical parameters (height, age,etc.) have shown the profound importance in speaker profiling applications and voice forensics using the speech data.

Researchers explored different features like Mel frequency cepstral coefficients, Linear prediction cepstral coefficients, fundamental frequency, formants, prosodic features, etc., are used for predicting physical parameters like height /age and other speaker characteristics. The extracted features are given to regression schemes like support vector regression, ANNs, Random forest, DNN models for predicting the physical parameter.

Speaker profiling is a challenging application area (Tanner and Tanner (2004)). In many cases, there is no control over the amount of available speech data from the target speaker. Therefore, such systems are required to provide accurate predictions using a minimum amount of speech data. Thus, development of speaker profiling methods in short duration audio is important.

The organization of the chapter is as follows, a brief insight about the literature on the physiological cues present in speech is detailed in Section 2.1. The different approaches in extracting the physical parameters like height, age, weight and other characteristics are detailed as prior art in speaker profiling in Section 2.2. The limitations and challenges which motivated to take up this work are briefed in Section 2.3. Finally the summary of the literature is presented in Section 2.4.

2.1 Physiological cues in speech

Literature shows that the physical dimensions of the speech production system are affected by the body build of a person. In general, a tall, well-built individual has lengthy vocal tract and large vocal folds (Layer and Truddgill (1979)). The previous studies on the predicted height and weight of a person and their correlations with the acoustic features like fundamental frequency (F_0), vocal tract length (VTL) have generated mixed results (Gonzalez (2003), Van Dommelen and Moxness (1995), Collins (2000)). The correlation values of 0.53 (male) and 0.57 (female) are reported between actual and perceived height values (Van Dommelen and Moxness (1995)). The previous studies have also reported that VTL estimated from the speech has only a weak correlation with body height (Necioglu *et al.* (2000), Pisanski *et al.* (2014)). The only exception is a study (Fitch and Giedd (1999)) involving people in the age group of 2.8 years to 25 years. This study reported the correlations between actual vocal tract length and height using magnetic resonance imaging (MRI). It shows that there is a strong correlation between vocal tract length and height of the speaker for the subjects considered(0.88 for children, 0.83 for female and 0.86 for male). It is also worthwhile noting that the sample size in this study for adult subjects (17 to 25 years) was quite small (six female and 13 male). Fitch and Giedd (1999), study shows that at the age of puberty, there is a substantial change in the vocal tract morphology. The gender difference becomes more significant at peri and post-puberty age. The studies also showed that formant frequencies are also closely related to vocal tract morphology, which helps in giving an acoustic cue to body size (Fitch and Giedd (1999)).

One of the speech cues associated with the body size dimension of the speaker is formant frequencies. They are weakly related to the body size dimensions such as height and weight, and chest circumference (Rendall *et al.* (2005), Evans *et al.* (2006), Greisbach (2007)). The voice characteristics of speech such as speech rate, sound pressure level, fundamental frequency, etc. are affected by the speaker's age (Müller (2006), Schötz (2007), Schötz and Müller (2007)). Other speech characteristics like harmonics Li *et al.* (2013), jitter (micro variations in fundamental frequency), shimmer (micro-variations of amplitude in frequency) occurs from age-related glottis deterioration (Müller and Burkhardt (2007), van Heerden *et al.* (2010)) of the speaker. These features contain information about speaker age. F0 will remain stable until the menopause (around 50 years) for female, when a drop happens, followed by either rise, fall or no change. Male F0 will drop until around the middle age when a rise follows until the old age (Schötz and Müller (2007)). With advancements in age, the following changes are observed in both the genders. Speech rate decreases, sound pressure levels (SPL) rate increases or remains relatively stable. F0 decreases till the menopause (around 50years) and then remain the same after the menopause for female speakers. In the case of males, F0 decreases slightly till 50 years and then increases (Schötz and Müller (2007)). Jitter and shimmer remained relatively stable for both male and female speakers. The shimmer is relatively remained stable for the female speaker from a young age to old age whereas for male speakers, increased slightly till the age group of 40 then decrease gradually till old age (Schötz and Müller (2007)).

Previous attempts by (Layer and Truddgill (1979), Van Dommelen and Moxness (1995)) in predicting the weight of a speaker, found a significant correlation to exist between weight and vocal fold traits like dimensions and mass. F_0 is significantly influenced by the obese and overweight people than normal persons. The obese and overweight people have lower F_0 values than the normal people (Souza and Santos (2018)). A few studies show that the listeners are able to perceive the weight (correlation of 0.724 for male and 0.627 for female speakers) and body build (Van Dommelen and Moxness (1995), Lass and Brown (1978), Lass *et al.* (1982)). By considering the listener's judgment, the weight of a speaker was identified, and the obtained correlation of 0.11, 0.14 and 0.09 for male, female and all speakers (Krauss *et al.* (2002)). Another study reports the correlation between log VTL and log weight as 0.862, 0.875

and 0.903 for children, females and males respectively (Fitch and Giedd (1999)). While a weak correlation exists between the weight of the speaker and the formant structure (Rendall *et al.* (2005), González (2004)), the speaking rate was found to be a useful feature used by human listeners in weight attribute estimation (Van Dommelen and Moxness (1995)).

2.2 Speaker Profiling Literature

While there is information about accent, height/age in the speech signal, the extraction of these parameters is challenging, as these parameters are also affected by numerous other factors such as the content being spoken, emotion and mood of the speaker, gender of the speaker etc. These factors degrade the performance of the height and age estimation methods.

2.2.1 Height Estimation

The height of a speaker can be estimated by standard sound specific features such as formants, F_0, sub-glottal resonances (SGR), short term spectral features and accumulated statistical features of the speech features across the sentence as input to the system.

The researchers had predicted the height of a speaker using the speech based features by using the short term features – Mel Frequency Cepstral Coefficients (MFCC) (Dusan (2005), Pellom and Hansen (1997)), Linear Prediction Coefficients(LPC)(Dusan (2005)), formant frequencies (Dusan (2005), Williams and Hansen (2013), Hansen *et al.* (2015)), sub-glottal resonances (Arsikere *et al.* (2012, 2013*a*)) and fundamental frequency (Dusan (2005)). Short term features like (MFCC, LPC) and formants of phone specific (vowels like /iy/,/ae/,/ey/,/ih/,/eh/ etc.) have shown a correlation of around 0.75. Similarly, a correlation of 0.59 has been observed for F_0 in estimating the height (Dusan (2005)). In an alternate approach (Arsikere *et al.* (2011)), the sub-glottal resonances are used for height estimation. SGRs are the resonance frequencies of sub-glottal (below the glottis) input impedance measurements from the top of the trachea. The SGRs are measured using the bark scale difference of the formants (Arsikere *et al.* (2013*a*)). These resonances are shown to be correlated with the height information, and a simple polynomial relation can then be employed to estimate the height. Using the SGRs, the overall mean absolute error (MAE) of 5.4 cm, root mean square error (RMSE) of 6.8 cm at the sentence level and 5.3 cm, 6.6 cm of MAE and RMSE respectively at speaker level on TIMIT data. However, this method requires a dataset where both the speech and glottal resonances were recorded in a parallel setting. Here the authors (Arsikere *et al.* (2013*a*)) have assumed that sub-glottal system acoustic length is proportional to height, observed the correlation between sub-glottal resonances and height by modeling system.

A few other studies use the vowel regions (/aa/,/ae/,/ao/,/iy/) to predict the height of a person by formant track regression (Hansen *et al.* (2015), Williams and Hansen (2013)). This method obtained the MAE is reduced to 6.36cm for male and 6.8cm for female speakers by considering a subset of speakers and selected sentences from TIMIT dataset. By fusing the formant track regression with height distribution based classification, the MAE is 5.37cm and 5.49cm for male and female speakers respectively. Later line spectral frequencies were added to the feature set resulting in a lower MAE 4.93cm and 4.76cm for male and female speakers respectively. However, these approaches require speech transcription and phone level alignment.

Another set of approaches that do not depend on the speech transcriptions use accumulated statistics of the short term speech features as input. These features are typically used on a regression scheme (Support Vector Regression (SVR), Artificial Neural Networks (ANN), etc.) in predicting the height of a person. For example, various statistics like mean, median, percentiles etc., are extracted from the short-term spectral features for automatic height estimation (Mporas and Ganchev (2009), Ganchev *et al.* (2010*a*)). Here a set of features are selected from a large pool of statistical features. A feature selection algorithm precedes the support vector regression which provides the estimate of the height and obtains MAE of 5.3cm and RMSE of 6.8cm on TIMIT dataset. A similar approach uses i-vectors (dimension reduced version of background Gaussian Mixture Model (GMM) statistics) followed by regression schemes (SVR, ANN, etc.) to estimate the height of a speaker (Poorjam *et al.* (2015, 2014)).

In another approach, the height is divided into different bins and the height class of the speaker is estimated (Pellom and Hansen (1997), Arsikere *et al.* (2013*b*)). For example the MFCC features are modelled using a background GMM to estimate the height class of a speaker (i.e., for a given utterance the height class is estimated). This approach using the TIMIT dataset reports results with a RMSE of 6.4 cm and 5.7cm for male and female speakers respectively (Arsikere *et al.* (2013*b*)).

Singh *et al.* (2016*b*) reports that the MAE performance of the default predictor (average value of that parameter over the training set) is often better than the results in literature such as (Williams and Hansen (2013), Mporas and Ganchev (2009), Ganchev *et al.* (2010*a*)). This study focuses on a bag of words representation instead of GMMs. The short term spectral features at multiple temporal resolutions are used to form a bag of words representation. For the TIMIT dataset, the MAE is 5.0 cm and RMSE is of 6.7 cm for male speakers and for female speakers the MAE is 5.0 cm and 6.1 cm RMSE. This study uses the short durations of speech data to estimate the height of a speaker (Singh *et al.* (2016*b*)).

2.2.2 Age Estimation

The accumulated statistics of the prosodic features and short term features can be used to estimate the age of the speaker. A popular approach is to use prosodic features such as jitter / shimmer, harmonics to noise ratio, fundamental frequency (Müller (2006), Müller and Burkhardt (2007), van Heerden *et al.* (2010)). These feature statistics are used by machine learning models like Artificial Neural Networks (ANN), Support Vector Machines (SVM), k-Nearest Neighbor (KNN) etc. to classify the age group of a speaker. By considering both male and female genders the age class accuracy is 94.61% using an ANN model in proprietary dataset (Müller (2006)). There have also been attempts to combine information from various levels such as short-term spectrum, prosodic features etc. These different feature sets are used to find the statistics of a background GMM. This statistics are used as a feature in SVM for the age classification task (Li *et al.* (2013), van Heerden *et al.* (2010)). With Interspeech2010 Para linguistic challenge dataset, the unweighted accuracy was 52% and weighted accuracy was 49.5% for the age classification problem (Li *et al.* (2013)). However, these efforts do not estimate the age, but only classify the input speaker as belonging to one of the age groups (e.g., kid, young adult, adult, etc.).

The statistical approaches adapted by researchers for age-group classification are Gaussian Mixture Model (GMM) Universal Background Model (UBM) (Müller and Burkhardt (2007), Metze *et al.* (2007), Bocklet *et al.* (2010)), support vector machines (Spiegl *et al.* (2009), Bahari *et al.* (2012), Li *et al.* (2010)), ANN (Poorjam *et al.* (2014)). These are followed by the statistical representation of short term features like MFCC, LPC, Perceptual Linear Prediction (PLP) coefficients, Temporal Patterns

(TRAPS) (Bocklet *et al.* (2010)) etc. In another approach, the age of a speaker is estimated by using a bag of words representation in place of background GMM from short-term cepstral features. In this work, short duration of speech data was considered and obtained MAE of 5.5 years & RMSE of 7.8 years for male, and for female speakers, MAE is 6.5 years & RMSE is 8.9 years on TIMIT dataset (Singh *et al.* (2016*b*)). Using the UBM based approach, the short-term features are represented as supervectors / i-vectors and these are used as input features to a classifier (Bahari *et al.* (2012), Sadjadi *et al.* (2016), Shivakumar *et al.* (2014)). Using NIST SRE08 and SRE10 data, the fusion of different short term features and i-vectors results in MAE of 4.7 years for male with correlation of 0.89, female MAE is 4.7 years with correlation of 0.91 (Sadjadi *et al.* (2016)). A more recent approach using the deep neural networks on the short utterances of telephone speech using long short term memory (LSTM) recurrent neural networks (RNN) (Zazo *et al.* (2018)) MAE and and correlation of male and female speakers are 8.72y, 0.37,and 7.95y, 0.54 respectively when 3s of speech is considered. An end to end deep neural network architecture using the x-vectors has also reported recently. Using only x-vectors on end to end system the MAE, correlations for 5s chunks of speech data are 5.78y, 0.74 for male, 4.23y, 0.87 for female respectively (Ghahremani *et al.* (2018)). Table 2.1 shows the summary of the prior works methods and features for height and age estimation tasks.

2.2.3 Body Build and other Characteristics Estimation

There are very few studies to estimate the other parameters like weight, shoulder size, chest circumference, shoulder to hip ratio, smoking habits, etc.,

The body size parameters like weight, neck etc. are predicted using F_0 and formants of all the vowels. The correlation between F_0 and first four formants with weight is 0.3 for male speakers (Rendall *et al.* (2005)). Another study (Evans *et al.* (2006)) shows the correlations of average fundamental frequency with shoulder circumference ($r = -0.29$), chest circumference ($r = -0.28$), shoulder-hip ratio ($r = -0.49$) and weight with formants is ($r = -0.43$).

Using the i-vector frame work weight is estimated and obtained the correlation of 0.56 for male and 0.41 for female speakers. The smoking habits are also predicted by using the i-vector framework with a log-likelihood ratio cost of 0.81 (Poorjam *et al.* (2014)).

Table 2.1: Summary of prior work in age and height estimation.

Reference	Motivation	Features	Model
Literature summary on Age			
Müller (2006), Müller and Burkhardt (2007), van Heerden *et al.* (2010), Metze *et al.* (2007)	Target advertisements	Pitch, jitter, shimmer, MFCC, LPC, etc.	ANN / SVM / GMM and fusion
Sadjadi *et al.* (2016), Bahari *et al.* (2012), Shivakumar *et al.* (2014)	Forensics, target advertisements	i-vectors	SVM / SVR
Ghahremani *et al.* (2018), Zazo *et al.* (2018)	Forensics, target advertisements, commercial applications	i-vectors/ x-vectors	DNN
Li *et al.* (2013), Bocklet *et al.* (2010), Li *et al.* (2010)	Target advertisements	MFCC, Prosodic features, Formants, Pitch, PLPs, TRAPs	SVM/ GMM .
Literature summary on Height			
Poorjam *et al.* (2015)	Forensics, biometric applications	i-vectors	LSSVR/ANN

Literature summary – Continued on next page

Table 2.1 – Literature summary – Continued from previous page

Reference	Motivation	Features	Model
Mporas and Ganchev (2009), Ganchev *et al.* (2010*a*)	Forensics, biometric applications	OpenSmile	SVR
Hansen *et al.* (2015), Pellom and Hansen (1997), Williams and Hansen (2013)	Forensic, biometric applications	LSF, Formants, MFCC	Linear Regression, GMM
Arsikere *et al.* (2012, 2013*a*, 2011, 2013*b*)	Relation between SGR and height	SGR	GMM, polynomial regression
Literature summary on Height and Age			
Poorjam *et al.* (2014)	Forensics, target advertisements	i-vectors	LSSVR/ANN
Singh *et al.* (2016*b*)	Forensics, target advertisements	Short term spectral features	Random Forest

2.3 Motivations & Challenges

While the past studies generate mixed results about the information present in speech relating to speaker height, body dimensions and age, engineering applications to extract these physical traits from speech have shown practically useful results (for example Hansen *et al.* (2015), Sadjadi *et al.* (2016)). However, in the existing literature, most of the significant results have focused on the estimation of height and age from long speech segments of few minutes (Sadjadi *et al.* (2016)) or by using hand labeled phoneme level features Hansen *et al.* (2015). The prior work on short duration speech shows that dealing with utterances of 5sec. length is challenging yielding significantly worse results making the systems inoperable for realistic applications Ghahremani *et al.* (2018).

Majority of the speaker profiling works of the past concentrate on estimating only two physical parameters age and/or height. The best results in height estimation are obtained by using features that are phoneme specific (Hansen *et al.* (2015), Dusan (2005), Williams and Hansen (2013)). This comes with the constraint on the system to have accurate transcription of the speech utterances with phone level alignment. The approaches involving SGR features Arsikere *et al.* (2012, 2013*a*, 2011, 2013*b*) require a separate dataset to learn the relationship between speech formants and the sub-glottal resonances. Other literature, often report the results on longer speech utterances using NIST recordings ($> 10s$) (Poorjam *et al.* (2015), Sadjadi *et al.* (2016), Ghahremani *et al.* (2018), Bahari *et al.* (2012), Shivakumar *et al.* (2014), Zazo *et al.* (2018)) and does not address speaker profiling from short utterances. Even for the i-vector based systems, the i-vectors may not be well estimated for short utterances (Sadjadi *et al.* (2016), Bahari *et al.* (2012), Shivakumar *et al.* (2014)). Also often gender specific speaker profiling results are not reported (Dusan (2005), Ganchev *et al.* (2010*a*)) and it was later reported that the gender-wise results of these methods are inferior to default predictor based on the mean of the training data performance genderwise (Singh *et al.* (2016*b*)). So far the only work that addressed both height and age estimation from short duration speech is Singh *et al.* (2016*b*). However, the prior work on short duration speech shows that dealing with utterances of $<$ 5sec. of speech in physical parameter estimation is challenging.

To the best of our knowledge, existing literature does not address the following and motivated to carry out this work,

1. The best height estimation system (Hansen *et al.* (2015)) uses the phoneme level transcription, which is practically tough to obtain in the speaker profiling applications.

2. The literature addresses height and/or age estimations. Other physical parameters like shoulder size and weight of a speaker are not explored.

3. Most of the works in the literature are monolingual (English) in estimating the physical parameters (mostly, height and age estimations only).

4. Many systems require at least utterances about 5-10s duration for physical parameter estimation. However, this may not be practical for forensic like scenarios.

2.4 Summary

As a summary of the literature survey, anatomical studies showed that a tall, well-built individual has lengthy vocal tract and large vocal folds, also from the scans of MRI shown the correlation between the VTL and height of a speaker. Fundamental frequency, formants, jitter, and shimmer affect the age of a speaker. Fundamental frequency influences the obese and overweight people than average persons.

The literature has shown that most of the studies that are carried out in physical parameter estimation are using monolingual speech data (English). Features like MFCC, LPC, Fundamental frequency, formants, SGR, short term spectral features, a large set of *OpenSmile* features are used in height estimation. Using the phoneme specific formant tracking regressions achieves the best height estimation system. Features based on SGRs need a separate measuring setup for sub-glottal resonances.

In the case of age estimation, most of the works carried out in the literature are estimating the age group of a person. The accumulated statistics of the prosodic features and short-term features can be used to estimate the speaker's age. The majority of the recent works in age estimation have used the large dataset of telephonic speech recordings SRE08 and SRE10. The age estimations are carried out using the utterance level representations using i-vectors and x-vectors. The recent works used x-vectors for an end to end estimation of the age of a speaker. The LSTM and RNN based DNN systems are also used to estimate the speaker's age from speech data.

To the best of our knowledge, there are no works carried out in estimating the multiple physical parameters in a multi-lingual setting using a common set of features. Similarly, there are no prior attempts on a joint prediction of multiple physical parameters.

Chapter 3

Speaker Profiling Features

Different features have been explored for speaker profiling in the literature. Mel frequency cepstral coefficients, Linear prediction cepstral coefficients, fundamental frequency, formants, prosodic features, etc., are used for predicting physical parameters like height /age and other speaker characteristics. However, there is no consensus for the best feature, while most of the above features result in comparable performance in estimating the physical parameters from the speech data.

In this chapter, we aim to come up with a common feature set for all the physical parameter prediction systems. The proposed features are extracted at the utterance level. These features do not require phone level transcriptions. We have explored different kinds of features that uncover the underlying spectral structure of the speech signal to estimate the physical parameters. The short-term mel spectrogram captures the gross level spectral characteristics used in predicting height and age of a speaker (Poorjam *et al.* (2015), Schuller *et al.* (2013), van Heerden *et al.* (2010)). The fundamental and formant frequencies contain information about physical parameters of a speaker (Hansen *et al.* (2015), Arsikere *et al.* (2013*a*)). The narrowband spectral harmonics capture the fine spectral structure on a coarse temporal scale. The log harmonics have used in estimating the age and gender of a speaker (Li *et al.* (2013)). Both frequency and amplitude of the spectral peaks have used as harmonic features (to capture jitter and shimmer characteristics of speech).

All the experiments are performed on the TIMIT dataset to estimate physical parameters using all these different types of features. The duration analysis has performed to determine the minimum amount of speech duration needed to estimate the physical parameters.

Estimating the physical parameters from the short duration of the speech data will help in many applications like targeted advertisements (predicting the age and height of a person will enable to specific requirements of the customer), forensic applications (in narrow down the suspects of hoax/threatening calls from a large number of suspects), speaker recognition, and speaker profiling tasks, etc.

The highlights of this chapter can be summarized as follows:

1. We have explored different features (MFCC, formants, harmonics) from speech data that do not require the phoneme level transcriptions to predict the physical parameters of a speaker.

2. We explored the multi-resolution features (short term spectrum –MFCC features, wideband spectrum – formants, narrow band spectrum – harmonic features) for physical parameter estimation system.

3. We proposed a common set of features (MFCC, formants and harmonics) for estimating the different physical traits of a speaker. The estimation errors from these different feature streams are complementary, which allows the combination of estimates from these feature streams to improve the results further.

4. The duration analysis of the proposed scheme shows that the state of the art results can be achieved using only around $1 - 2$ seconds of speech data.

5. To the best of our knowledge, this is the first attempt to use a common set of features for estimating the different physical traits of a speaker from a short duration of the speech.

The rest of the chapter is organized as follows. Section 3.1 details about the dataset used in this chapter. Section 3.2 describe the block diagram of physical parameter estimation system. Section 3.3 to Section 3.5 provides the details about the feature extraction and also the representation these features at sentence level. The prediction model and evaluation metrics are detailed in Section 3.6 and Section 3.7 respectively. Section 3.8 details about the experiments conducted on the TIMIT dataset. The key findings and the summary of the chapter are briefed in Section 3.9.

3.1 Dataset

The TIMIT dataset has 630 speakers (Garofolo *et al.* (1993)) and each speaker has contributed ten recordings. Each utterance is considered as a separate input data sample. TIMIT dataset has the details of the height and age of a speaker. The height values in the dataset range from 145 cm to 204 cm and speakers' ages range from 21 years to 76 years. Each input utterance has an average of $1 - 3$ seconds of speech data. The height and age details are considered for estimations in the experiments conducted in this chapter. The statistics details of the height and age of the dataset are given in Table 3.1. The statistics of height are shown in Figure 3.1 and age are shown in Figure 3.2.

Table 3.1: Statistics of each parameter in the TIMIT dataset (Garofolo *et al.* (1993))

Physical Characteristic	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
Male Speakers				
Height (cm)	157.48	203.20	179.73	7.09
Age (y)	21	76	30.52	7.57
Female Speakers				
Height (cm)	144.78	182.88	165.80	6.71
Age (y)	21	67	30.03	8.70
Male and Female Speakers				
Height (cm)	144.78	203.20	175.50	9.47
Age (y)	21	76	30.37	7.98

3.2 System Overview

The Physical Parameter Estimation System flow is shown in the block diagram in Figure.3.3. In the pre-process stage, we remove the silence from the speech data using Voice Activity Detection (VAD) algorithm (Tan and Lindberg (2010)). We analyze the speech signal at different resolutions in the spectral domain and explore the possibility of predicting the speaker characteristics from the same set of features.

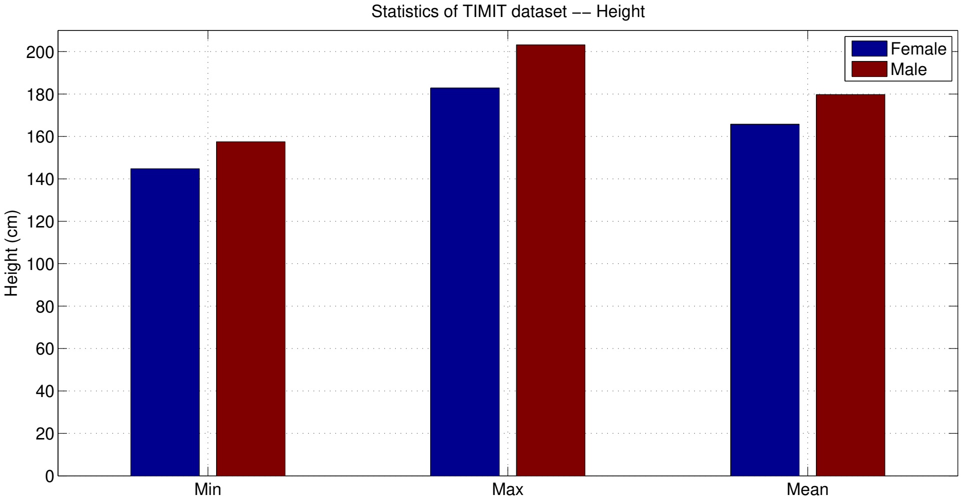

Figure 3.1: Statistics of TIMIT dataset – Height

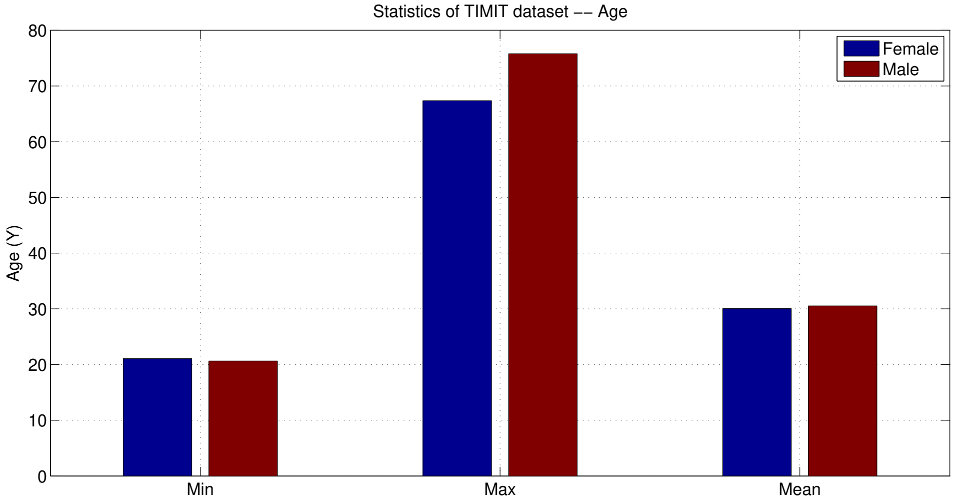

Figure 3.2: Statistics of TIMIT dataset – Age

From the literature, it is noted that formants, fundamental frequency, harmonics, and short term cepstral features contain information pertaining to physical parameter estimation. We extract a different set of features which uncover the underlying spectral structure of the from the speech signal. The short-term mel spectral features capture the gross level spectral characteristics. The formant frequencies represent the resonant frequencies in the speech signal. The narrowband spectral harmonics captures the fine spectral structures on a coarse temporal scale. Utterance level statistics of these features is represented as a single vector for every speech file separately. These features do not require phone level transcriptions. Finally, these are fed as input to the Support Vector Regression (SVR) to predict the speaker's desired physical parameters.

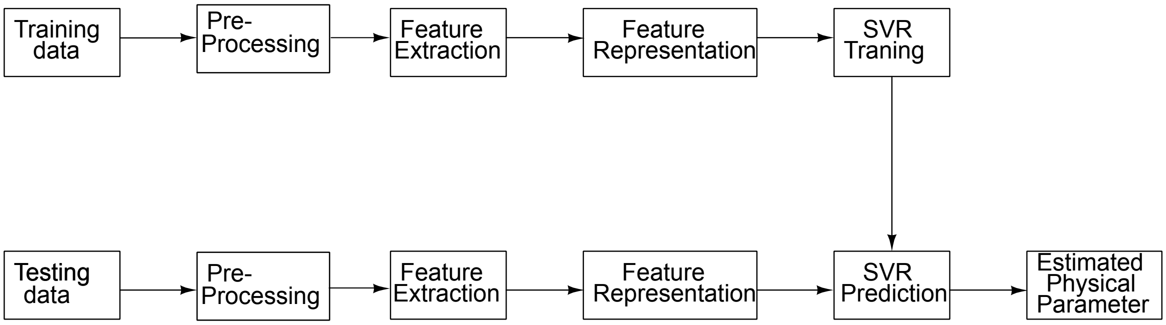

Figure 3.3: Block diagram of physical parameter estimation system

3.3 Spectral Features Extraction

We have explored different short term spectral features Mel Frequency Cepstral Coefficients and Mel Filter Bank features. In order to normalize the linguistic effect, the extracted frame level features are represented as super-vectors which are similar to $i-vectors$ in speaker verification and speaker identification task (Reynolds (2002), Campbell *et al.* (2006)).

3.3.1 Cepstral Features

The Mel Frequency Cepstral Coefficients (MFCC) features are the most commonly representations used in speaker recognition. The MFCC features are have some information relating to the vocal tract length (Müller and Burkhardt (2007), Dusan (2005)). In the past, the MFCC features and their statistics have been employed followed by the regression scheme for height and age estimation (Li *et al.* (2013), Mporas and Ganchev (2009), Ganchev *et al.* (2010*a*), Poorjam *et al.* (2014)).

In our work, we extract mel frequency cepstral coefficients and mel filter bank features from the speech signal. 20 mel frequency cepstral coefficients (using a window length of 25 ms with a shift of 10 ms) are extracted along with delta and double delta features (yielding 60 MFCC features). We also use the logarithm of the mel spectral energy in short-term windows ($25ms$ with a shift of $10ms$) of the speech signal. The mel filter bank features are the short energy features computed prior to the Discrete Cosine Transform (DCT) in the MFCC feature computation. We extract 40 mel filter bank features. The short spectral features contain the phonetic information as well as the speaker information. However, these features are modified significantly by speech content in terms of phoneme variability. We adopt an approach similar to supervector (Reynolds (2002)), which can summarize the gross spectral changes in order to normalize the effect of phonetic information in the short-term spectral representation.

3.3.2 Sentence Level Representation

To find the sentence level representations, we use the statistics of the background model components. In order to form a background UBM model, a Gaussian Mixture Model (GMM) is estimated from short-term spectral features. Let $\mathbf{x}_i$ and $\mathbf{y}_i$ be input MFCC feature (i.e, $\mathbf{x}_i \in \mathcal{R}^{60}$) and mel-filter bank feature (i.e, $\mathbf{y}_i \in \mathcal{R}^{40}$) corresponding to frame i respectively. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ represents the input MFCC feature vectors and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T\}$ represent mel filter bank features for an input utterance with T frames. The diagonal covariance GMM -UBM is trained on MFCC features. The GMM probability density is :

$$f_{UBM}(\mathbf{x}) = \sum_{j=1}^{M} w_j \mathcal{N}(\mathbf{x}; \boldsymbol{\mu_j}, \mathbf{C}_j) \tag{3.1}$$

where $\mathbf{x}$, denotes input feature vector (MFCC) and $\boldsymbol{\mu}_j, \mathbf{C}_j$ represent the mean and the diagonal covariance matrix of the j^{th} GMM component with weight w_j respectively. The frame level first order statistics for a given frame i and each GMM component j is computed as:

$$\mathbf{f}_i^j = \mathbf{y}_i p(j|\mathbf{x}_i), \tag{3.2}$$

where the a-posteriori probabilities of a GMM component j is given by:

$$p(j|\mathbf{x}_i) = \frac{w_j \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}_j, \mathbf{C}_j)}{\sum_{j=1}^{M} w_j \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}_j, \mathbf{C}_j)}. \tag{3.3}$$

We then concatenate all $\mathbf{f}_i^j$ for all GMM components to obtain a super vector $\mathbf{F}_i = [f_i^1, f_i^2, \ldots, f_i^j, \ldots, f_i^M]$ which represents the utterance. The first order statistics for a given utterance is:

$$\mathbf{F} = \frac{1}{T}\sum_{i=1}^{T}\mathbf{F}_i \tag{3.4}$$

Intuitively, if each GMM component j corresponds to a different sound class, the average of $\mathbf{f_j^i}$ over the frames i would represent the short-term spectral average of frames that belong to that sound class. Therefore the size of the first order statistics per utterance is $j \times length\ of\ feature\ vector$ (here $j = 256$ components and length of mel filter bank feature is 40, i.e, $256 \times 40 = 10240$). These statistical features are fed to support vector regression to estimate the physical parameter.

3.4 Fundamental Frequency and Formant Features

The fundamental and formant frequencies represent the fundamental and resonant frequencies in the speech signal. These features have shown to be influenced by speaker's height and age (Krauss *et al.* (2002), Li *et al.* (2013)). We compute these features as follows.

We compute the fundamental frequency from a wideband analysis of speech signal (temporal window size of $20ms$ with a shift of $10ms$). The estimation is performed with the PEFAC algorithm (Gonzalez and Brookes (2014)) which combines noise rejection and normalization while ensuring temporal continuity in the estimates using dynamic programming. For physical parameter estimation, we use the statistics (mean, standard deviation and percentiles) of the time varying fundamental frequency computed over the given speech recording.

The formant frequencies are estimated by picking the peaks of an auto regressive (AR) model of the power spectrum. The peaks of the wide-band (window length of $20ms$ with a shift of $10ms$) spectrum can approximately represent the formant structure. We use an AR model of order 18 to extract peak locations results in nine peak locations. The first four peak locations are used to capture formant frequencies (denoted as F_1, F_2, F_3 and F_4). The wide-band spectrogram of speech for a vowel regions and the corresponding formant frequency trajectories are depicted in Figure.3.4.

The first four formant frequencies (F_1, F_2, F_3, F_4) are extracted from the speech signal. We analyze the correlation between the fundamental frequency (F_0) and the other formant frequencies with the height values. The studies have shown F_0 is inversely proportional to height of a speaker (indicating that the speakers with more height

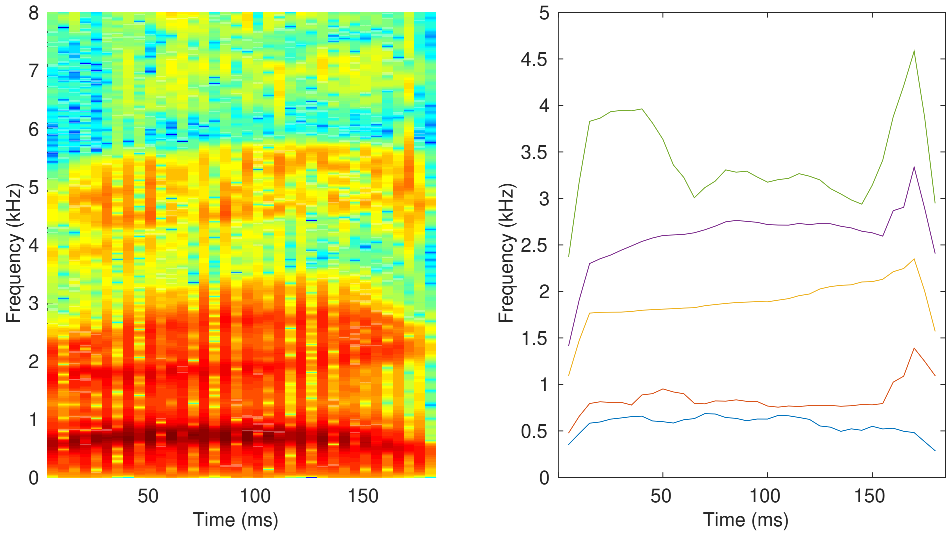

Figure 3.4: Spectrogram for vowel /AE/ and corresponding formants

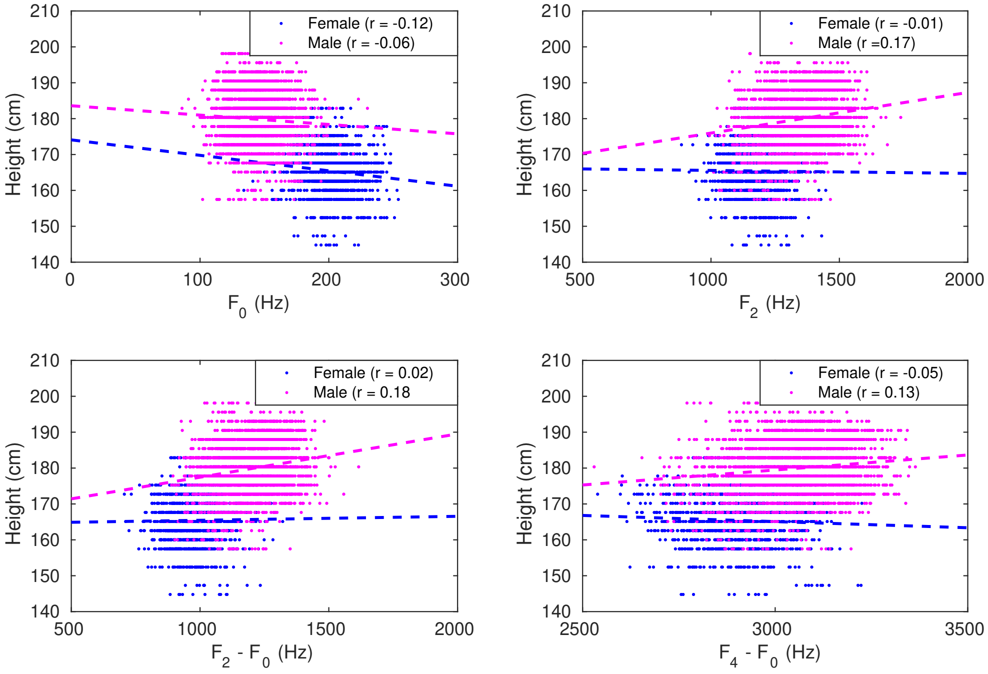

Figure 3.5: Scatter plot of fundamental and formant frequency estimates with the speaker height for TIMIT training set. Value in the brackets shows the correlation (r) between formants and corresponding physical parameter (height) for male and female speakers. The best fit line is also shown for both male and female speakers separately.

values have low fundamental frequency and vice-versa for speakers with lesser height values) (Van Dommelen and Moxness (1995), Evans *et al.* (2006), Greisbach (2007)). The fundamental frequency (F_0), has a weak correlation with height ($r = -0.12$) for female speakers. Similarly, for male speakers F_2 showed a weak correlation with height value ($r = -0.17$). The correlations of male height vs F_0 ($r = -0.06$) and female height vs F_2 ($r = -0.01$) are relatively insignificant. Literature has reported weak correlations between body build of the speaker and different functions of formant frequencies such as dispersion (Fitch (1997)), average formant position (Puts *et al.* (2012)), formant spacing (Reby and McComb (2003)), difference between F_0 and formants (Rendall *et al.* (2005)). For example, we find the correlations between difference of F_0 and formants ($F_1 - F_0$, $F_2 - F_0$, $F_3 - F_0$, $F_4 - F_0$), Figure.3.5 depicts some of the results for the training portion of TIMIT dataset. It is observed that, $F_2 - F_0$ and $F_4 - F_0$ have weak positive correlation for male speakers ($r = 0.18$ and $r = 0.13$ respectively) and weak correlations for female speakers with height values (Rendall *et al.* (2005)).

3.4.1 Sentence Level Representation

Speaker identification systems have used mean value of pitch, range of pitch etc., as utterance level features (Peskin *et al.* (2003)). In this work, we use a similar approach where each sentence is represented using statistics of the log fundamental frequency and log formant frequencies across the utterance. We use percentiles of log-peak locations in the short-term spectrum of speech (computed over time). The peak locations in the spectrum include the fundamental frequency and formant frequencies. In addition to the percentiles, the statistics of peak locations (in log-frequency scale) like the mean and standard deviation are used to estimate the physical parameters like height/age. These statistics can implicitly capture the average value, range and variance of fundamental frequency and formants.

3.5 Harmonic Features

The long-term features like jitter (micro variations in the fundamental frequency), shimmer (micro variations in the amplitude) carry the cues related to speaker characteristics (van Heerden *et al.* (2010)). In addition to the conventional mel frequency spectrum and formants, we also experimented with the use of harmonic structure of the speech signal. The harmonics are formed as a result of vocal fold vibration during voiced speech. We investigated in both amplitude and frequency locations of the long-term features in order to estimate the physical parameters from the speech data.

It was empirically observed that the height influences the harmonics. Figure.3.6 shows the narrowband spectrogram portion of the same TIMIT (*sa2 – Don't ask me to carry an oily rag like that*) utterance spoken by tall (height value of 175 cm) and short (height value of 152 cm) female speakers (left panel). We also highlight the magnitude response of a single speech frame on the right panel. We can see that for the same vowel /oy/ (from the word *oily*), the taller person has smaller harmonics and vice-versa. The linguistic content for the chosen speech frame has been verified to be the same. The distance between two successive harmonics is also listed. As seen here, the harmonics are more closer for the taller speaker. It was empirically observed that the height influences the harmonics. The taller person has smaller harmonics and vice-versa. It is also noted that the harmonics appear along with a jitter (not exactly the multiple of a fundamental frequency). While this would mean that the peak locations are not truly harmonic, we continue to refer to the peaks in the narrowband spectrum as harmonic frequencies. It has been shown that variations in frequency (jitter) and amplitude (shimmer) contain useful information about age

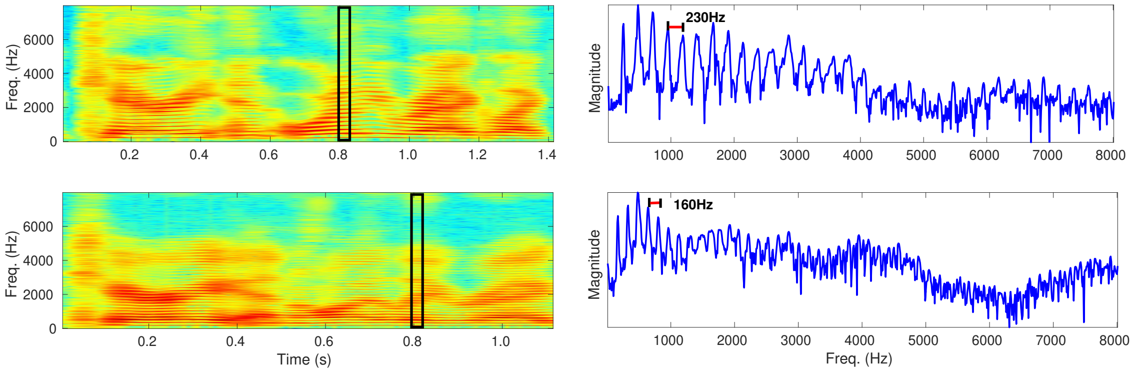

Figure 3.6: Illustration of relation between harmonic frequencies and speaker height. These plots are computed for same underlying *sa*2 speech utterance for two female speakers with height of 152 cm (top panel) and 175 cm (bottom panel). The left panel is the narrowband spectrogram and the right panel is the magnitude spectrum of the frame highlighted in the left. The distance between the harmonic frequency estimates is listed in the right panel. The taller person has smaller harmonics as compared to a short person.

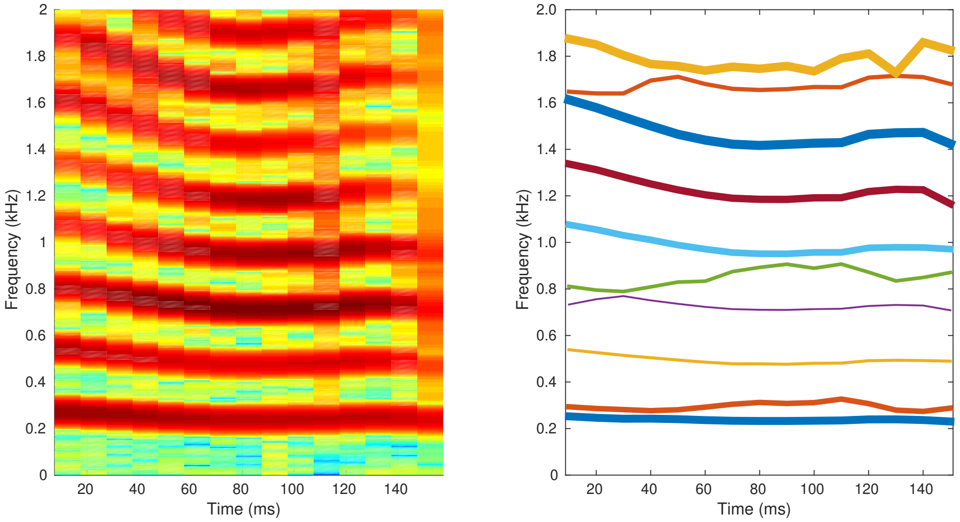

Figure 3.7: Spectrogram for vowel /AE/ and corresponding trajectories of first 10 peaks locations in a narrow-band spectrogram estimated using an AR model.

as well (Müller and Burkhardt (2007)).

The harmonic frequencies are estimated as the peak locations of a higher order AR model. The logarithm of the frequency and amplitude of spectral peaks are computed at each frame. We use the similar approach which we used in Section 3.4.1 to represent the entire utterance as a sentence level feature. Each sentence is represented by the percentiles of log frequency and log amplitude values of spectral peaks over the utterance. The percentiles of harmonic frequencies represents the mean range and jitter in the harmonics. Similarly, the statistics on amplitude can contain shimmer in addition to average and range values. The collection of these statistics is referred to as "harmonic features" in this work. Figure.3.7 shows a short term spectrogram of the speech along with estimated harmonics.

The scatter plot for first harmonic frequency percentiles (25 and 50) on TIMIT training data are shown in Figure.3.8 for both male and female speakers. It is observed that there is a weak negative correlation in case of height and age for percentiles 25 and 50 for both male and female speakers. We also observe that the log magnitude statistics (percentiles) of the first two harmonic frequencies show a weak negative correlation with both age and height for both male and female speakers. These statistical harmonic features are used as input for support vector regression algorithm. The

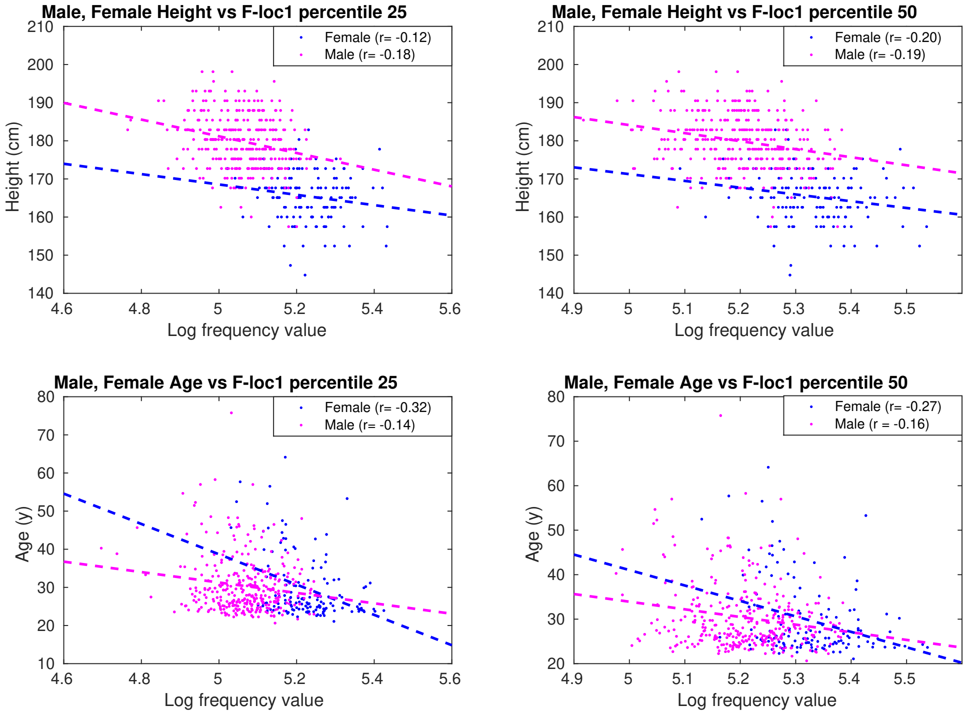

Figure 3.8: Scatter plot of Harmonic percentiles (25 and 50) vs physical parameter (height and age) for male and female speakers of TIMIT training data. Correlation (r) value between harmonic percentile and physical parameters (height and Age) is given in brackets for male and female speakers. The best fit line is also shown for both male and female speakers separately.

frequency location features capture jitter features and amplitude features captures shimmer features.

3.6 Support Vector Regression

Different linear and non-linear regression models have been experimented with in the context of physical parameter prediction (Dusan (2005), Ganchev *et al.* (2010*b*), Arsikere *et al.* (2014)). In this work, support vector regression (Smola and Schölkopf (2004)) is used as the model for predicting the target of each physical parameter values given the statistically represented features explained previous sections. Let us denote the set of pair of input features along with target values as $\{(\mathbf{y_1}, t_1), (\mathbf{y_2}, t_2), \ldots (\mathbf{y_m}, t_m)\}$. The function $f(\mathbf{y}) = \mathbf{w^T y} + b$ corresponds to the linear SVR to learn and performs the following optimization:

$$\begin{aligned} \min \frac{1}{2}\mathbf{w^T w} \text{ subject to} \\ |\mathbf{w^T y_i} + b - t_i| < \epsilon \end{aligned} \tag{3.5}$$

where, b is the bias term and the "fit" function is controlled by the parameter ϵ. The maximum deviation from the target values is ϵ. The optimization of the SVR objective function can be solely carried out in terms of the dot product of the data points among themselves. The SVR optimization function aims to reduce the deviation from the target values by the parameter ϵ.

3.7 Evaluation Metrics

The common metric used by researchers in measuring the error are Root Mean Square Error, Mean Absolute Error and Correlation. We use the same metric to measure the error in all our experiments.

Root Mean Square Error (RMSE):

RMSE is the most commonly measuring metric used to measure the difference between predicted values by a model and actual targets values observed. This is given by the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_{tar,i} - x_{pred,i})^2}{N}} \tag{3.6}$$

where x_{tar} are the target values and x_{pred} are the predicted values of each utterance i.

Mean Absolute Error (MAE):

MAE measures how far the predicted values are away from the observed target values. This is given by the following equation:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_{tar,i} - x_{pred,i}| \tag{3.7}$$

where x_{tar} are the target values and x_{pred} are the predicted values of each utterance i.

3.8 Experiments and Results

All the experiments in this chapter are performed on the TIMIT dataset. The standard Train and Test splits of the dataset are used for the experiments performed to estimate the height and age of a speaker. We have 462 speakers (326 male and 136 female speakers) for the training set and 168 speakers for the test set (56 female and 112 male speakers). The training and validation splits have 4610 utterances, including 3260 utterances from male speakers and 1360 utterances from female speakers. The test split has 1120 utterances from male speakers and 560 utterances from the female speakers. We experimented with a different feature set like Mel filter bank first-order statistics, formants, and harmonics to predict the physical parameters. Finally, we combine the individual feature predictions to get a better estimate.

3.8.1 Target Mean Predictor

Target Mean Predictor (TMP) is the prediction of mean of test samples by blindly predicting the mean of training data without looking at the features. It gives the best estimate in the absence of speech information. Here for our case we took the mean of each physical trait of training speech samples without looking into speech information and predicted the error of each sample of the test speech samples using this mean. The TMP values of height and age are tabulated in Table 3.2.

Table 3.2: The MAE and RMSE values of target mean predictor on TIMIT dataset.

	Male		Female		All	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Height	5.3	7.0	5.2	6.5	7.4	9.0
Age	5.7	8.1	6.4	9.2	5.9	8.4

3.8.2 Individual Feature Results

In order to understand the effect of each feature separately, we evaluated the individual performance of the features. All hyper parameters of the system (e.g., kernel choice for SVR) and the order of the models were fixed based on the validation dataset performance.

We first perform a speech activity detection (Tan and Lindberg (2010)) and then extract the speech features. In order to extract the first order statistics (Fstats), we first train a 256 component GMM with 60 dimension MFCC features ($\mathbf{x}_i$). The Fstats are computed with 40 dimensional mel filter bank features ($\mathbf{y}_i$) using the Eq. 3.4. This gives $40 * 256 = 10240$ dimensional vector. The Fstats are fed to a support vector regression model to predict the physical parameters. A linear kernel is used for the support vector regression.

Fundamental frequency and formant features are extracted by picking the resonant frequencies of an all-pole model. A 18^{th} order (fixed based on validation set) model is used with a $20ms$ length window with $10ms$ shift. The 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentile values across the entire utterance are employed as features. A linear kernel is used in the SVR.

A similar approach was followed in case of harmonic features. Thirty harmonics were extracted from an 80 order all-pole model, computed over a longer time window (length $60ms$ and shift $10ms$). The same set of percentiles are computed and used as input to a SVR with a third degree polynomial kernel (the order, window size and kernel are fixed based on the validation dataset). We separately evaluate the performance of harmonic frequencies, amplitudes as well as both together.

For comparison purposes, we also compute the Training data Mean Predictor (TMP). This just corresponds to providing the sample mean of the training data targets (physical parameters) as the estimate for any input, i.e., without using any evidence from the test speech. Figure.3.9 illustrates the performance of each feature as well as the TMP. In addition to the Fstats, and formants features, the figure also illustrates the effect of estimated harmonic frequency locations (F-loc) and corresponding amplitudes (Amp) as well as their combination ('harmonic' features). Both formants and Fstats have shown minimal improvement over TMP for both the genders in estimating the height of a speaker. The harmonic features show improvements only for female height and age estimation. In both these cases, the combination of harmonic features performs better than using either frequency locations or amplitudes. The performance improvement over TMP MAE is of 2.71% when Fstats are used for predicting height of male speakers. Similarly, for female speakers the improvement in MAE is of 4.01%, 3.23% , and 3.13% when formants, Fstats and harmonics are used respectively. For in predicting the age, all the features have shown a better performance when compared with TMP MAE for both the genders. For the male speakers,

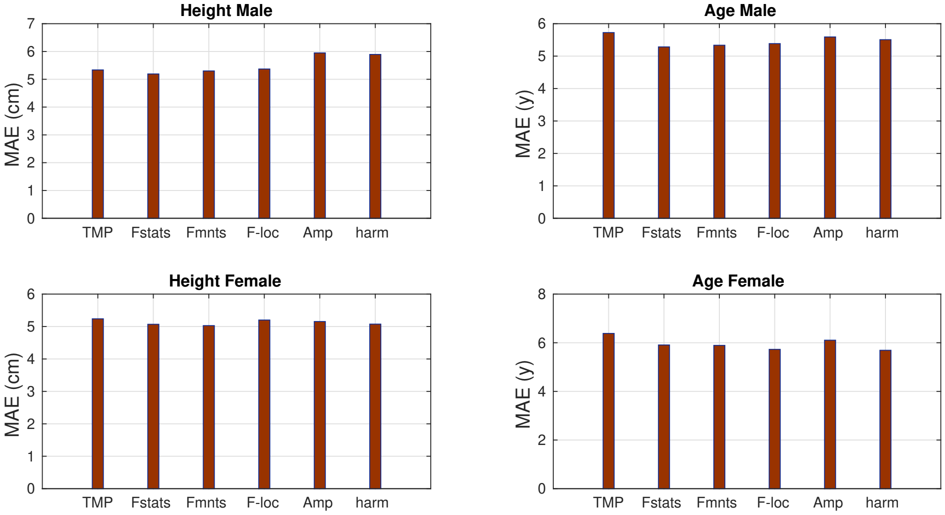

Figure 3.9: Mean absolute error comparison with target mean predictor (TMP) and prediction of different systems using first order statistics (Fstats), formants (Fmnts), harmonic frequency locations (F-loc), amplitude (Amp) and harmonic features (harmonic frequency locations & amplitude features together: harm) for height (left side) and age (right side) estimation using the TIMIT dataset.

the improvement in MAE is of 6.8%, 3.82% and 7.7% for formants, harmonics and Fstats respectively. Similarly, for female speakers the improvement in MAE is of 7.71% 10.85% and 7.38% when formants, harmonics and Fstats respectively.

In short, for both the male and female speakers, all the features have shown better performance compared with TMP MAE in age estimation, and Fstats and formants showed better MAE in height estimation. Harmonics shows better MAE in female speakers' height estimation but not in the male speaker's height estimation.

3.8.3 Feature Combination Results

Since the features contain diverse information about the speech data, we investigate the level of system complementarity in terms of generating uncorrelated errors. In our analysis, we found that the different feature sets produce different height and age estimation errors for a large number of validation speakers. With this knowledge, we attempt a simple averaging of the individual regression outputs to improve the final height and age estimates. We have made three different sets of feature combinations

Table 3.3: Comparison of the proposed feature combinations – Comb -1 (Fstats + formant + frequency locations), Comb -2 (Fstats + formant + amplitude), Comb -3 (Fstats + formant + harmonic features (amplitude + frequency locations)) with state-of-the-art results on TIMIT dataset.

Height (cm) Estimation						
	Male		Female		All	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
TMP	5.3	7.0	5.2	6.5	7.4	9.0
Ganchev *et al.* (2010*a*)	-	-	-	-	5.3	6.8
Arsikere *et al.* (2013*a*)	5.6	6.9	5.0	6.4	5.4	6.8
Singh *et al.* (2016*b*)	**5.0**	**6.7**	5.0	6.1	-	-
Comb-1	5.2	6.8	5.0	6.3	5.2	6.8
Comb-2	5.2	6.9	4.8	6.2	5.2	6.7
Comb-3	5.2	6.8	**4.8**	**6.1**	**5.2**	**6.7**

Age(y) Estimation						
	Male		Female		All	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
TMP	5.7	8.1	6.4	9.2	5.9	**8.4**
Singh *et al.* (2016*b*)	5.5	**7.8**	6.5	8.9	-	-
Comb-1	5.3	8.2	5.8	9.2	5.5	8.7
Comb-2	5.3	8.2	5.6	8.8	5.4	8.6
Comb-3	**5.2**	8.1	**5.6**	**8.7**	**5.4**	8.5

of Fstats and formant features with either harmonic frequency location (Comb -1) or amplitude (Comb -2) or harmonic features (both frequency and amplitude features: Comb -3). Table 3.3 reports the results along with the recent baseline(Singh *et al.* (2016*b*)).

The relative improvement of height prediction MAE for Comb-3 w.r.t TMP is 1.89% and 8.33% for male and female speakers respectively. Similarly, the relative improvement of age prediction MAE is 8.77%, and 14.29% for male and female speakers respectively. In case of RMSE, the relative improvement in height prediction of Comb-3 w.r.t to TMP is 2.94% and 6.15% for male and female speakers respectively. Similarly, for age prediction there is an 5.75% relative improvement for female speakers and no improvement for the male speakers.

We performed a paired t-test comparing the absolute errors from proposed system (Comb -3) and the default predictor (TMP) in a gender-wise manner. For both the tasks of height and age estimation, the proposed system is significantly different from the TMP ($p < 0.05$) across both the gender cases.

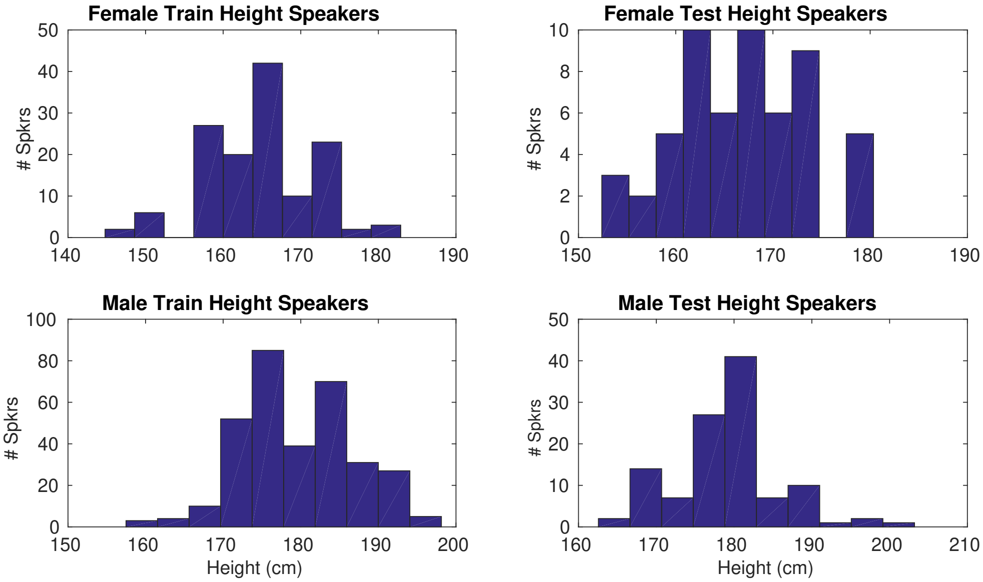

Figure 3.10: Speaker Height – Training data (Left) and Test data (Right)

In case of height estimation, we also compare with three other baselines. The error metrics MAE and RMSE of the proposed systems as well as the baseline results are presented in Table 3.3. In case of female speakers both MAE and RMSE performances of Comb -3 are better than the baseline for height estimation. In order to gain further insight into the proposed height estimation system, we analyze the performance of height and age estimation of the data in different subgroups of Comb -3.

Table 3.4 lists various subgroups along with the height estimation performance and number of training speakers in each subgroup. It can be seen that large errors occur for speakers in the sub groups which are at the two extreme height values (row 3 & 6 for male speakers and 2 & 5 for female speakers) in Table 3.4. This may be due to the small amount of training data available for these groups. The gender specific histogram of speaker heights for both training and testing datasets are depicted in Figure.3.10. We also observe that there is a mismatch in train and test height histograms. Such mismatches could have also resulted large error in extreme values of height.

In case of age estimation, the only work that has reported results on short segments

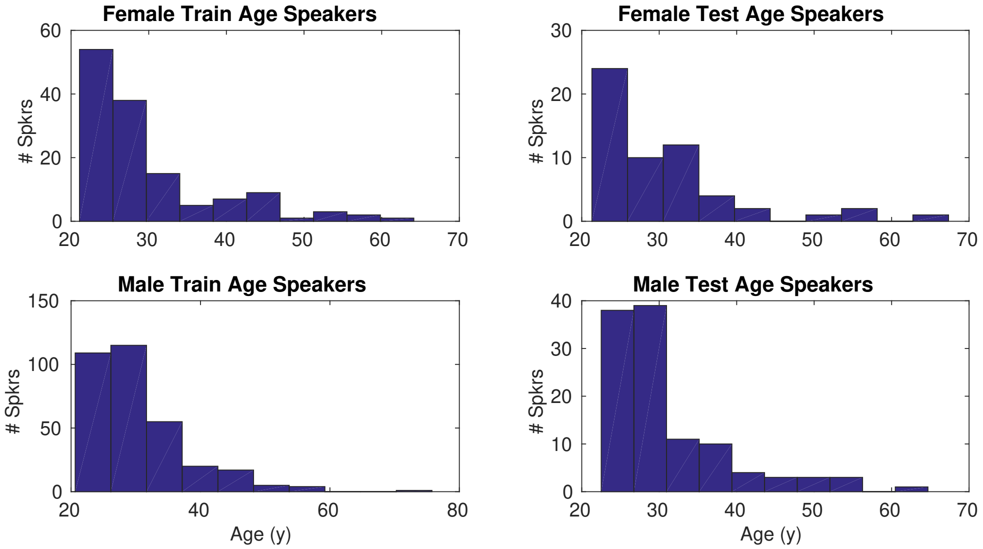

Figure 3.11: Speaker Age – Training data (Left) and Test data (Right)

Table 3.4: Height (h) estimation errors (MAE and RMSE in centimeters(cm)) across different height subgroups using TIMIT test data

		Male			Female		
Sl. No.	Range	# Train Spkrs	MAE	RMSE	# Train Spkrs	MAE	RMSE
1.	$145 \leq h < 150$	0	-	-	2	-	-
2.	$150 \leq h < 160$	2	-	-	20	9.3	9.6
3.	$160 \leq h < 170$	15	11.9	12.2	75	2.5	3.0
4.	$170 \leq h < 180$	137	4.7	5.7	35	6.4	7.1
5.	$180 \leq h < 190$	140	2.9	3.7	3	14.9	14.9
6.	$190 \leq h < 203$	32	12.5	13.1	0	-	-

in TIMIT is by Singh *et al.* (2016*b*). Comparison of this baseline with our results and TMP is presented in Table 3.3. Note that in case of female speakers the baseline had a higher MAE as compared to TMP. The proposed systems outperforms the baseline results and TMP in terms of MAE for male and female speakers. However, RMSE value is at par with TMP in case of Comb -3 male speakers and better than state of the art in female speakers in all the feature combinations. We analyzed the performance of Comb -3 for age estimation system by dividing the data into different subgroups as

Table 3.5: Age (a) estimation error (MAE and RMSE in years) across different age subgroups using TIMIT test data

		Male			Female		
Sl. No.	Range	# Train Spkrs	MAE	RMSE	# Train Spkrs	MAE	RMSE
1.	$20 \leq a < 25$	67	4.6	4.8	47	2.7	3.0
2.	$25 \leq a < 30$	132	1.8	2.1	46	2.0	2.4
3.	$30 \leq a < 35$	66	2.9	3.4	14	4.7	5.2
4.	$35 \leq a < 40$	28	7.8	8.1	9	8.8	8.9
5.	$40 \leq a < 45$	13	13.0	13.1	9	13.0	13.1
6.	$45 \leq a < 55$	16	22.2	22.4	7	24.9	25.0
7.	$55 \leq a < 65$	3	35.5	35.5	3	21.9	21.9
8.	$65 \leq a < 76$	1	-	-	0	35.0	35.1

shown in Table 3.5. The RMSE is high over the TMP is due the presence of last three age groups (from 45 years to 75 years) in both the genders (refer Table 3.5). All these age groups have very few training speakers. Therefore, the RMSE error in these three groups are large (greater than 22y) and is dominates the overall RMSE performance. The histogram of gender specific speaker age in both training and testing datasets are depicted in Figure.3.11. It can be seen that there are very few number of speakers above 45 years in training.

3.8.4 Duration Analysis

In order to analyse the minimum amount of speech required for the task, we try to evaluate the performance of the system at different utterance durations. We initially use the standard TIMIT database and evaluated the system for different time lengths of input speech ranging from 0.25s to full length. The mean absolute errors for these different lengths of speech were compared with TMP with height and age of a speaker and shown in Figure.3.12.

We performed a genderwise paired t-test comparing the absolute errors from proposed system (Comb -3) and the default predictor (TMP) for different durations of speech data. We find that (with criterion of $p < 0.05$) the proposed approach results in significant improvements in age estimation for all durations considered (starting from 0.5sec.) for both the genders and the relative improvement in MAE is 3.15% for males and 15.84% for female speakers. In the case of height estimation, the proposed approach results in significant improvements starting from 1.5 sec. duration of audio segments and the relative improvement in MAE for male speakers is 2.87% and for female speakers is 5.58%. Also, as the duration of the available speech increases, the MAE reduces as expected. Subsequently, when sufficient amount of speech data is available, the mean absolute error get saturated.

It can be noted that even with roughly 1s of speech data, when both male and females speakers are considered, the model is able to obtain prediction error MAE of 5.27cm at par with Ganchev *et al.* (2010*a*) in speaker height prediction. As the available speech duration increases, this prediction error saturates around 5.2 cm when both genders are considered. Similarly for age prediction when both male and female speakers are considered together, the minimum duration of speech required to get the state-of-the-art prediction error is 0.5s (i.e, 5.5 years MAE). Even with around 3s speech available, the prediction error is marginally better (5.41 years). Gender wise results on duration analysis are also shown in Figure.3.12. About, 2s of speech data is required to get a performance comparable to the full length data.

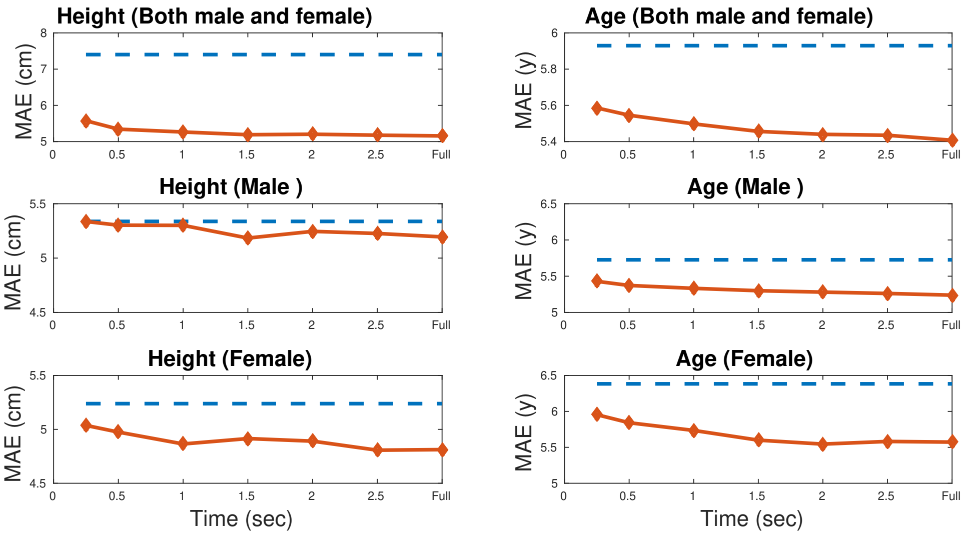

Figure 3.12: MAE vs duration of utterance, for physical parameters' (Height, Age) estimation from TIMIT database. The horizontal dashed line represent target mean predictor (TMP) benchmark.

3.9 Summary

In this chapter we have explored three different sets of multi-resolution features – Mel spectrum first-order statistics, statistics of formants, and harmonics for speaker profiling task.

All the features showed a performance improvement over the training mean predictor in age estimation. Whereas, formants and Fstats showed an improvement in the height of both male and female speakers over TMP. In the case of height estimation, male speakers showed an improvement of 2.71% in MAE, and for female speakers, the improvement in MAE is of 4.01%, 3.23%, and 3.13% when formants, Fstats, harmonics have used respectively. In the case of age estimation, the improvement in MAE w.rt TMP for male speakers is 6.8%, 3.82%, and 7.7% for formants, harmonics, and Fstats, respectively. Similarly, for female speakers, the improvement in MAE is 7.77%, 10.85%, and 7.38% when formants, harmonics, and Fstats, respectively.

Furthermore, these individual features are shown to be complementary, and a simple averaging improves the performance by achieving an MAE of 5.2 cm for male and all (male and female) and 4.8 cm for female speakers in height estimation. For age estimation, the MAE is 5.2 years, 5.6 years, and 5.4 years for males, females, and all speakers using the TIMIT dataset.

The duration analysis reveals that the prediction error of each physical parameter of a speaker is less than the training data mean predictor with as little speech as 0.5s. Also, with around $1 - 2$ seconds of data, the MAE obtained is as good as the state-of-the-art results, which were achieved using the full duration of the audio signal ($> 10s$).

Chapter 4

Estimation of Multiple Physical Parameters

Existing speech corpora have limited information about speaker metadata. Most of them have either physical characteristics or accent information, but often not about both. For example, the most common dataset TIMIT (Garofolo *et al.* (1993)) has only age, height, and gender information about the speakers. There is no information about other physical parameters or about the accent. The popular Speaker Recognition Evaluation (SRE) challenge datasets (NIST-SRE, Martin and Greenberg (2009, 2010)) have the information about smoking habits and native country. They don't have linguistic information. Other datasets such as 2010 Interspeech Paralinguistic Challenge(ComParE) dataset (Schuller *et al.* (2013)), Fisher English Corpus (Cieri *et al.* (2004)), SpeechDat II dataset (GermanSpeechDat (II)) provides only the gender and age group information of the speaker. The CMU Kids(Maxine Eskenazi) dataset only contains the grade information of the kids. None of these datasets provide any details about physical parameters beyond height and age. The only exception to this is the Copycat corpus (Lehman and Singh (2016)) that has details of height, weight, and age, but the speakers are limited to children. Similarly, there are also datasets that provide the accent information of the speakers, such as Accents of British Isles (ABI-1) corpus (DArcy *et al.* (2004)) and the CSLU-Foreign Accent English (FAE) (Lander) datasets. There is a need for a dataset with richer metadata in this context, including the linguistic content for speaker profiling systems.

Another limitation of current datasets is that most of the available datasets are monolingual (English). On the other hand, multilingual data available (for example, the Babel dataset (Harper (2013))) do not have detailed speaker profiling information.

We describe our efforts in collecting multilingual, multi-accent datasets from different Indian languages across different states of India. In this regard, to perform the speaker profiling task in a multilingual environment, we have collected two different datasets. They are called;

1. Audio Forensic Dataset (AFDS).
2. NITK-IISc Multilingual Multi-accent Speaker Profiling (NISP) dataset.

The first dataset (AFDS) is collected for a pilot study in speaker profiling. It contains 207 speakers across 12 different native Indian languages[1] along with English. The average number of speakers per native language is 17 (maximum number of speakers is Hindi(68), and minimum is for Urdu(1)). The number of utterances per speaker are four for each recorded language.

Later, a bigger dataset NISP dataset with 345 speakers is collected from five Indian languages[2] as well as English. NISP dataset has around 60 speakers per native language (Hindi has 103 speakers). The number of utterances per speaker are around 40-50 for each recorded language.

The highlights of the chapter are summarized as follows

1. Two different multi-lingual and multi accent datasets in Indian languages have collected.

2. Extended the physical parameter estimation system to other physical parameters like shoulder size, waist size, and weight along with age, and height of a speaker from the speech data.

3. Single set of multi-resolution features are used for estimating the all the physical parameters.

4. Estimated the physical parameter estimation system accuracy on multi-lingual setting for both the collected datasets.

5. Performed the duration analysis on the multi-lingual setting of physical parameter estimation system to know the least amount of speech data required to estimate the physical parameter of the speaker.

[1] Bengali, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Oriya, Tamil, Telugu, Urdu

[2] Hindi, Kannada, Malayalam, Tamil and Telugu

6. Performed the rigours analysis on the effect of language in estimating the multiple physical parameters when the model is trained with multiple languages and multiple accents.

The rest of the chapter is organized as follows. Section 4.1 details about the design of the datasets, metadata collection, recording environments, and protocol of speech data. The statistics of the collected datasets AFDS and NISP datasets are described in Section 4.2 and Section 4.3 respectively. Section 4.4 explains the potential applications of the collected datasets. Section 4.5 and Section 4.6 details about the experiments performed on the AFDS and NISP datasets. They describe the experiments conducted in the datasets and results. Different experiments are performed to assess the effect of duration and multilingual setting in speaker profiling. Finally, a summary of the chapter is briefed in Section 4.7.

4.1 Design of Datasets

We have collected two multilingual datasets for the physical parameters estimation task to perform the speaker profiling task in a multilingual environment from the short duration of speech data.

4.1.1 Metadata

The speakers who participated in contributing speech data for these databases consisted of students, academic staff, and faculty members of different educational institutions across southern India. Informed consent is obtained from the speakers to use the data for academic and research activities. The linguistic, regional, and physical traits are collected from each speaker, along with the speech data. The metadata information collected in these datasets is the following.

1. **Linguistic Information**

    (a) Native language (L1) of the speaker and whether the speaker can read text from L1.
    (b) Medium of Instruction. (noted if the speaker would have studied in local state language medium other than English medium).
    (c) Second language (L2) - Most commonly spoken language other than L1.

2. **Regional Information**

    (a) The geographic location of the native place (or the place where the subject has lived dominantly).
    (b) Current place of residence. (speakers' present residing district and state).

3. **Physical Characteristics Information**

    (a) Gender of the speaker (Male / Female).
    (b) Age in years
    (c) Height – without wearing sandals / shoes, measured in centimeter using a wall-mounted measuring tape.
    (d) Shoulder size – measured at the widest point of shoulders between acromion bone with the individual's arms at their side in centimeters using body measuring tape.
    (e) Waist size – measured as circumference above the hip in centimeters using the body measurement tape.
    (f) Weight – in kilograms using a standard digital weighing machine.

4.1.2 Speech Recording Environment

The audio recordings were collected in the environments like a normal classroom/seminar hall in each of the educational institution. All necessary precautions are taken care to avoid ambient noise and reverberations.

4.1.3 Speech Data and Recording Protocol

Both the datasets has different recording protocols and speech data. All the volunteers are asked to read the given text in their Native language (L1) as well as English at two different instances. Each dataset has no overlap among the speakers as well as provided text.

4.1.3.1 AFDS Recording Protocol

AFDS has twelve different native languages speech data (refer to Table 4.3) along with English. A continuous contextual text is provided to speakers from the daily news articles in both the native language and English, text is saved in UTF-8 format. Each volunteer is asked to read aloud sentences in both English and native Indian languages one after the other for each session, depending on the speaker's first language.

The recordings were made using a head-phone microphone (Logitech H110 stereo headset) at a 16 kHz sampling rate. All the speech recordings are made using *Audacity* software. All the entire dataset's speech recording is collected using the same head-phone microphone to avoid any channel variations across recordings. All speakers contributed roughly 2 minutes of data in 3 sessions each lasting 40 seconds.

4.1.3.2 NISP dataset Recording Protocol

The speech data was collected using a high-quality microphone (with Scarlett solo studio, CM25 a large-diaphragm condenser microphone). The data was sampled at 44.1 kHz with a bit-rate of 16 bits per sample. In order to avoid any channel variations across recordings, all the speech samples were collected using the same microphone device for this dataset creation.

The text data used in the reading task for the speakers were presented in the L1 language (refer to Table 4.5) as well as in English in two different sessions. The text provided to speakers was taken from the daily news articles as unique sentences without any contextual continuity from one sentence to another in both native and English texts. This setting was made to avoid any prosodic continuity in the reading task. Separately, a continuous short story section was given to the speakers in both the L1 and English languages to have contextual continuity effects in the reading task. Along with these sentences, we had also used five common sentences for every speaker. This includes two TIMIT *sa1* and *sa2* sentences and three general news article sentences in English language (to perform speaker profiling in text-dependent manner). Similarly two common sentences were also made in the native language text. Overall, each subject provided 20-25 unique sentences in L1 and English, 20-25 contextual sentences in L1 and English, 5 common sentences for English, and 2 sentences from L1. Each speaker was instructed to read aloud in a clear voice with a close-talking microphone.

The audio recording setup is made by using publicly available software, namely "*Speech Recorder*[3]" and with *Focusrite Scarlett solo studio* audio recording device by connecting it to a laptop. This audio recorder device has gain controller to adjust

[3] Speech Recorder software is freely available in the following address, *https://www.bas.uni-muenchen.de/forschung/Bas/software/speechrecorder/*

the gain and amplitude of the speech signal while recording. The software enables a graphical user interface (GUI) to display each sentence on the speaker's screen. It is monitored and controlled by a controller on another display. The participant is asked to read out the text aloud, displayed on the monitor in a comfortable sitting posture. The controller also verified the content, which is being read, in order to avoid any reading errors made by the speaker.

The statistics of collected datasets are detailed in the following sections.

4.2 Characteristics of Audio Forensics Dataset

We have collected this multilingual and multi-accent dataset from volunteers (students) of the National Institute of Technology Karnataka-Surathkal (NITK) to address the speaker profiling task's challenges. The volunteers are from different parts of the country. There are a total of 207 speakers includes 161 male and 46 female speakers. All the speakers fall in the age group of 18 to 35 years. Each speaker has contributed at least 120 seconds, and utmost 150 seconds of speech data in three sessions; each session lasts 40seconds. The set of speakers in the dataset is linguistically diverse, consisting of 12 different native tongues. The distribution of male and female speakers across India is shown in Table 4.1 and the same is displayed in Figure 4.1. From the speakers' entire distribution, 142 speakers (99 males and 43 females) are recorded in the seminar hall, and the remaining 65 speakers (62 males and 3 females) are recorded in the class room. The statistics of collected physical traits are tabulated in Table 4.2.

The distribution of speakers across the different native languages as well as gender-wise distribution, is shown in Table 4.3. English* in the table, indicates that these speakers can not read their native language; hence, only English speeches have recorded. For the rest of the speakers both native and English language speeches have recorded. The total number of utterances in this dataset are 1,489, out of which 1,161 are male speaker utterances, and 328 are female speaker utterances. The total number of native language utterances are 728 and there are 761 English utterances in the dataset. This dataset has a total of 7.92 hours of native language speech data and 8.43 hours of English speech data. The total number of utterances per language and total speech data duration of each recorded language are given in Table 4.4.

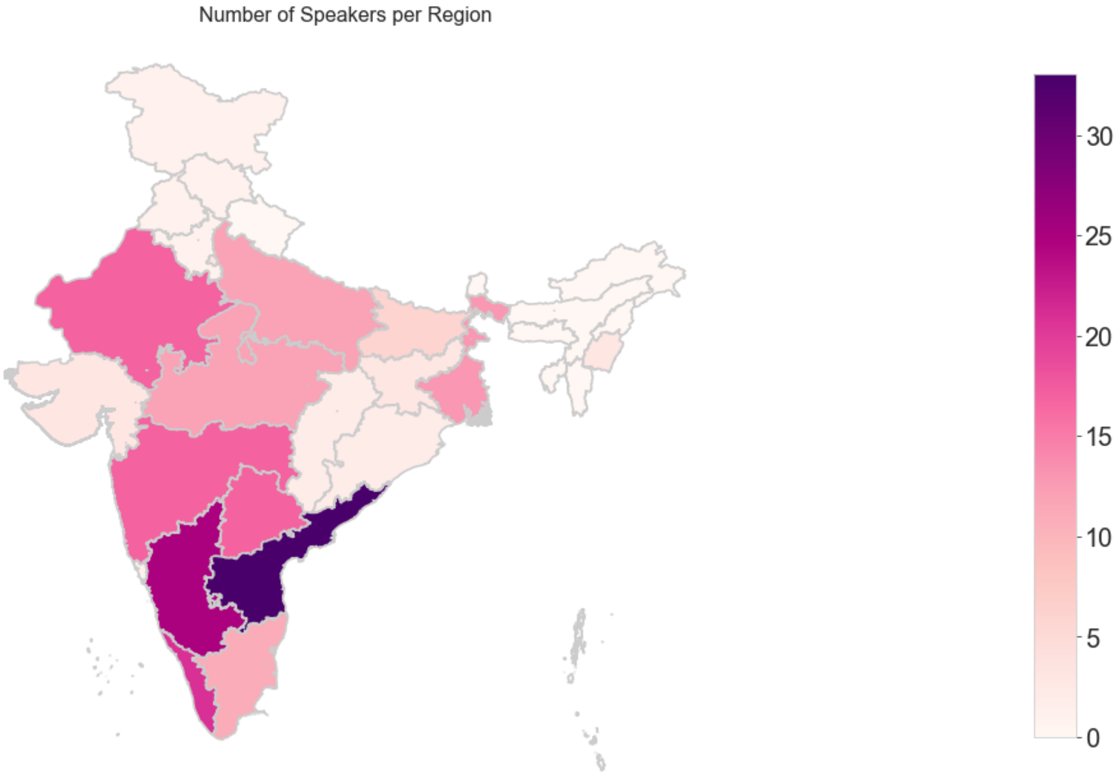

Figure 4.1: Native geographic region of the speakers in the AFDS dataset.

4.3 Characteristics of NITK-IISc Speaker Profiling Dataset

We attempt to overcome some of the available datasets' limitations by collecting multilingual, multi-accent datasets from five Indian native languages. This dataset is called *NITK-IISc Multilingual Multi-accent Speaker Profiling*[4] (NISP) dataset.

This dataset has collected from the volunteers of different colleges of south India, namely, Sree Vidyanikethan Engineering College, Tirupathi, Andhra Pradesh for Native language – *Telugu*, KSR College of Engineering, Tiruchengode, Tamilnadu, and NITK for Native language – *Tamil*, College of Engineering Thalassery, Kerala and NITK for Native language – *Malayalam*, NITK for Native language – *Kannada*, Indian Institute of Sciences (IISc), and NITK for Native Language – *Hindi*.

[4] This dataset is made publicly available in the following address, *https://github.com/iiscleap/NISP-Dataset.* This dataset is freely available for academic and research purposes with standard license agreements.

Table 4.1: Statewise distribution of male and female speakers in AFDS dataset

Sl.No	State	Male	Female	Total
1	Andhra Pradesh	28	5	33
2	Bihar	6	–	6
3	Chattisgarh	2	–	2
4	Delhi	2	1	3
5	Goa	1	–	1
6	Gujarat	3	–	3
7	Haryana	–	1	1
8	Himachal Pradesh	1	–	1
9	Jammu & Kashmir	1	–	1
10	Jharkhand	3	–	3
11	Karnataka	18	7	25
12	Kerala	11	10	21
13	Madhya Pradesh	7	5	12
14	Maharastra	15	2	17
15	Manipur	3	–	3
16	Odisha	2	–	2
17	Punjab	1	–	1
18	Rajasthan	16	1	17
19	Tamil Nadu	8	3	11
20	Telangana	16	1	17
21	Uttar Pradesh	8	3	11
22	West Bengal	7	6	13
23	Puducherry	1	–	1
24	Nepal	1	–	1
	TOTAL	161	46	207

The NISP dataset has 345 speakers, which includes 219 male and 126 female speakers. The dataset has five native Indian languages (namely Hindi, Kannada, Malayalam, Tamil and Telugu), as well as Indian, accented English. Each speaker provided around 4-5 minutes of speech data in each language. The distribution of speakers across the different native languages as well as gender-wise distribution, is

Table 4.2: Statistics of each parameter in the AFDS dataset (Kalluri *et al.* (2016))

Physical Characteristic	**Minimum**	**Maximum**	**Mean**	**Standard Deviation**
Male Speakers				
Height (cm)	156	188	171.0	6.7
Shoulder width (cm)	40	53	45.0	2.5
Waist size (cm)	68	112	86.0	7.6
Weight (kg)	45	107	67.9	11.1
Age (y)	18	37	23.32	3.10
Female Speakers				
Height (cm)	147	169	157.6	5.1
Shoulder width (cm)	30	45	38.4	2.6
Waist size (cm)	64	97	80.4	7.0
Weight (kg)	39	77	52.7	6.9
Age (y)	18	30	23.50	2.44
Male and Female Speakers				
Height (cm)	147	188	168.0	8.5
Shoulder width (cm)	30	53	43.5	3.7
Waist size (cm)	64	112	84.7	7.8
Weight (kg)	39	107	64.5	12.1
Age (y)	18	37	23.36	2.96

shown in Table 4.5. The total number of utterances in this dataset are $28,268$, out of which $17,844$ are male speaker utterances, and $10,424$ are female speaker utterances. The total number of native language utterances are $13,577$ and there are $14,691$ English utterances in the dataset. This dataset has a total of 24.83 hours of native language speech data and 32.03 hours of English speech data.

The total duration of speech in hours and the total number of utterances corresponding to each native language along with English speech are shown in Fig 4.3. The gender-wise statistics of each physical parameters are given in Table 4.6. The total number of speakers from each region per accent is shown in Fig 4.2.

Table 4.3: Distribution of native languages', male and female speakers of AFDS

Sl.No	language	Male	Female	total
1	Only English*	6	4	10
2	Hindi	56	12	68
3	Kannada	16	5	21
4	Malayalam	9	9	18
5	Manipuri	3	0	3
6	Marati	14	2	16
7	Tamil	9	2	11
8	Telugu	39	6	45
9	Bengali	4	6	10
10	Gujarathi	2	0	2
11	Odiya	2	0	2
12	Urdu	1	0	1
Total Speakers		161	46	207

Table 4.4: Speech duration in minutes (Dur) and number of utterances (Utt) for each language in AFDS

		Male		Female		Total	
Sl.No	Language	Dur	# Utt	Dur	# Utt	Dur	# Utt
1	English	395.8	590	109.9	171	505.7	761
2	Hindi	138.3	220	28.7	48	167.1	268
3	Kannada	38.3	51	12.1	17	50.3	68
4	Malayalam	21	27	20.2	29	41.3	56
5	Marathi	33.6	53	4.5	8	38.1	61
6	Tamil	22.3	35	4.8	8	27.1	43
7	Telugu	92.8	137	14.6	23	107.3	160
8	Bengali	9.5	16	14.6	24	24.1	40
9	Gujarati	5	8	0	0	5	8
10	Oriya	4.7	7	0	0	4.7	7
11	Urdu	2.9	5	0	0	2.9	5
12	Manipuri	7.2	12	0	0	7.2	12
Total		771.3	1161	209.4	328	980.7	1489

Table 4.5: Distribution of native languages', male and female speakers of NISP dataset

Sl.No.	Native Language	Males	Females	Total
1.	Hindi	76	27	103
2.	Kannada	33	27	60
3.	Malayalam	35	25	60
4.	Telugu	35	22	57
5.	Tamil	40	25	65
Total Speakers		219	126	345

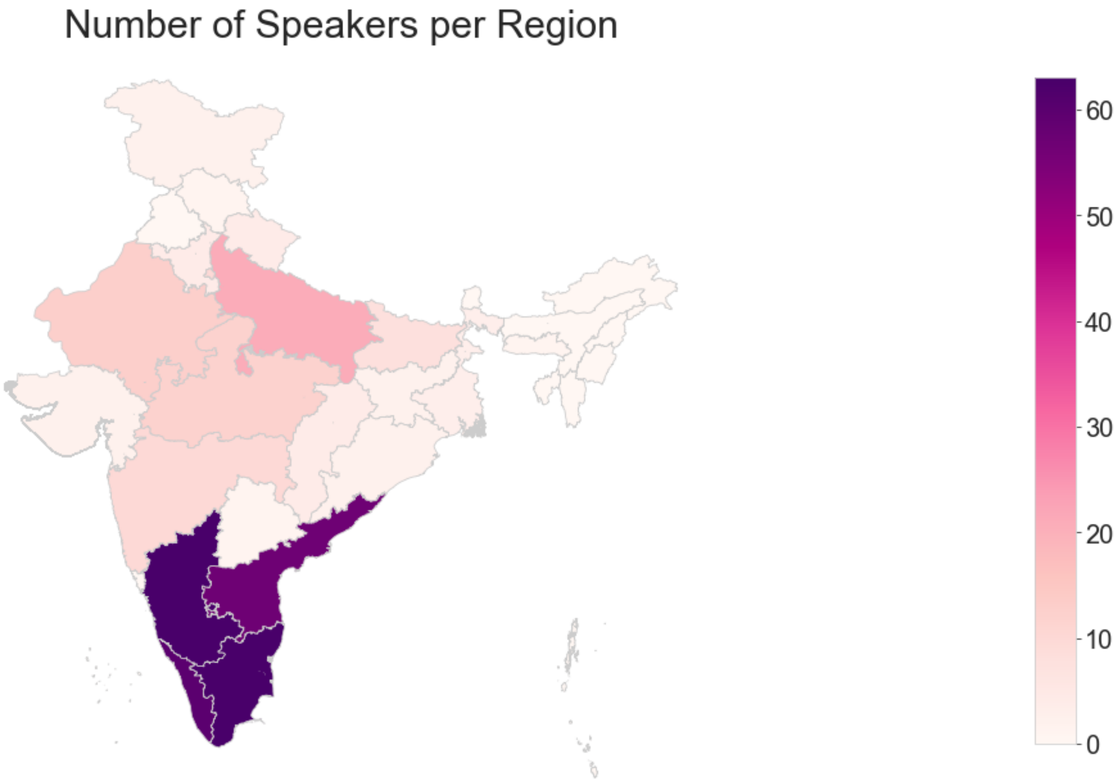

Figure 4.2: Native geographic region of the speakers in the NISP dataset.

4.4 Potential Applications

Both the datasets NISP and AFDS datasets provide a wide range of various applications depending on the task requirement. These datasets provides the ability to explore profiling applications in text dependent or independent fashion, accent/language identification experiments, speaker recognition as well as multilingual speech recognition

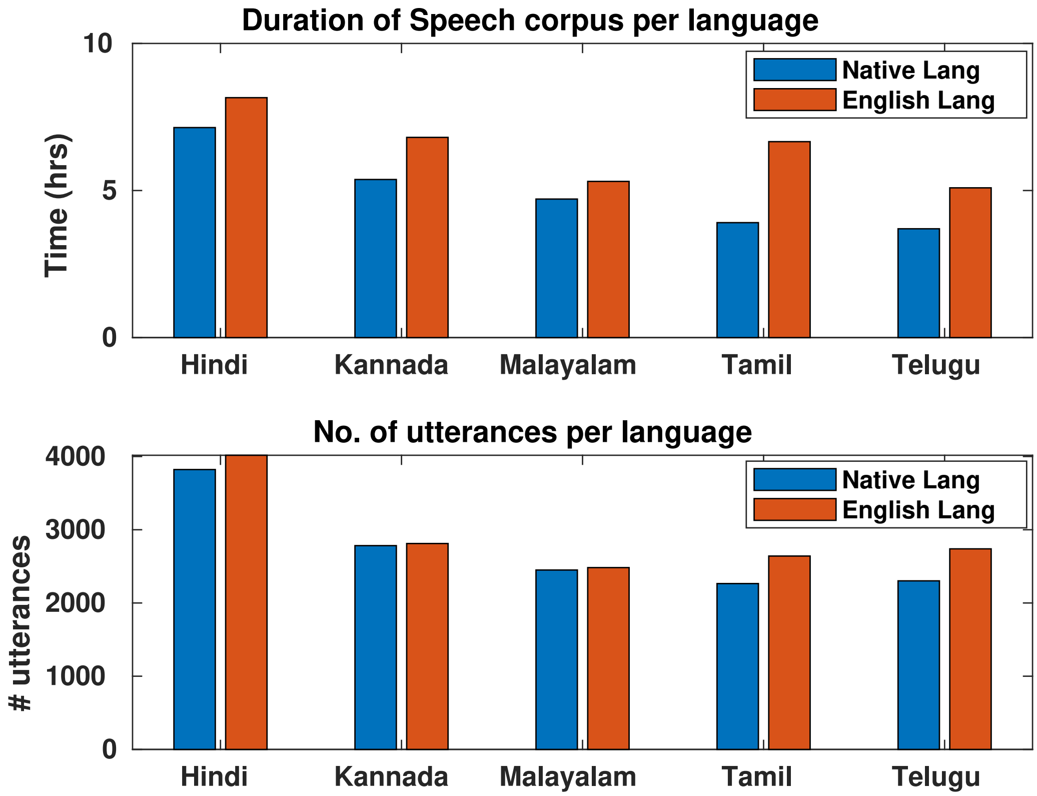

Figure 4.3: Number of utterances and speech duration of each language (both native language and English speech data) in the NISP dataset

experiments.

4.4.1 Accent & Language Identification

Identifying the accent and L1 of the speaker is an important cue in the voice forensic applications as well as in smart speaker and dialog systems. The NISP dataset enables research to explore accent related effects on speech. This database allows both L1 identification from L2 as well as language identification based on the 5 L1 languages.

4.4.2 Speaker Recognition

The NISP dataset, while being much smaller in scale, can be used to fine-tune the large neural network models with more multi-accent and multilingual variabilities. We hypothesize that this can improve the robustness of speaker recognition systems. In

Table 4.6: Statistics of each parameter in the NISP dataset

Physical Characteristic	Minimum	Maximum	Mean	Standard Deviation
		Male Speakers		
Height (cm)	151.0	191.0	171.6	6.7
Shoulder width (cm)	32.0	55.0	44.7	3.2
Weight (kg)	43.4	116.5	69.4	11.9
Age (y)	18.0	47.5	24.4	5.6
		Female Speakers		
Height (cm)	143.0	180.0	158.9	6.8
Shoulder width (cm)	30.0	53.0	39.7	3.4
Weight (kg)	34.1	86.2	56.5	10.5
Age (y)	18.3	46.5	25.1	6.1
		Male and Female Speakers		
Height (cm)	143.0	191.0	166.9	9.1
Shoulder width (cm)	30.0	55.0	42.9	4.0
Weight (kg)	34.1	116.5	64.7	13.0
Age (y)	18.0	47.5	24.7	5.8

addition, multilingual speaker verification with mismatched languages in enrollment and test data can be useful for bench-marking speaker verification systems.

4.4.3 Speech Recognition

This dataset has potentially rich text information in both English and all the native languages (Hindi, Kannada, Malayalam, Tamil and Telugu). All these transcription, after manual verification, are recorded in UTF-8 format.

4.5 Experimental Results on AFDS

Multiple physical parameters are estimated in a multilingual setting on AFDS using the same approach described in Chapter 3. The utterance level statistics are computed for each of the speakers.

In order to compute the first-order statistics on both AFDS and NISP datasets, 20 MFCCs along with deltas and double deltas are extracted with a window size of $25ms$ with a shift of $10ms$, together constitutes 60 features. And also, 40 filter bank features (window size of $25ms$ with a shift of $10ms$) are extracted separately. The GMM UBM learned from training data of TIMIT dataset is used, as the number of training speakers are less in these datasets (whereas TIMIT dataset has 630 speakers). The first-order statistics are computed on both AFDS and NISP datasets using the Eq.3.4 (refer to Section 3.3.2 in Chapter 3). The statistics are computed for formants, harmonic features over the entire utterance (explained in Section 3.4, Section 3.5 in Chapter 3) along with first-order statistics of the dataset are fed to the support vector regression (detailed in Section 3.6 in Chapter 3) separately for each physical parameter estimation.

For the evaluation purpose, the dataset is split into training and testing splits. Training data has 137 speakers consisting of 104 males and 33 female speakers. The training split has 951 utterances includes both English and native languages. Whereas for test split has 70 speakers with 57 males and 13 female speakers. The test split has 538 utterances consists of both English and native languages. Both the training and testing splits are linguistically diverse and proportional. The statistics of train and test splits are given in Table 4.7. The standard error metrics Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), are used to measure the actual and predicted targets' errors. The target mean predictor (TMP) is used to compare the error performance of the system.

4.5.1 Individual Feature Results

The statistics computed from formants, frequency locations, amplitude, harmonics and first order statistics from Mel filter bank features, are fed to SVR to train for each physical parameter. The SVR is trained in a multilingual setting on all different native languages data of the AFDS and tested as well.

The mean absolute error of each feature is compared with the target mean predictor of each physical parameter (height, shoulder size, waist size, weight and age) is shown in Figure.4.4. The Fstats, and formants shows better MAE performance for all the physical parameters when male speakers and both gender speakers are considered, but not with female speakers.

The Fstats have showed an improvement in MAE when compared with TMP for height,shoulder size, waist size, weight and age is 25%, 29.5%, 1.8%, 15.1%, and 3.9%

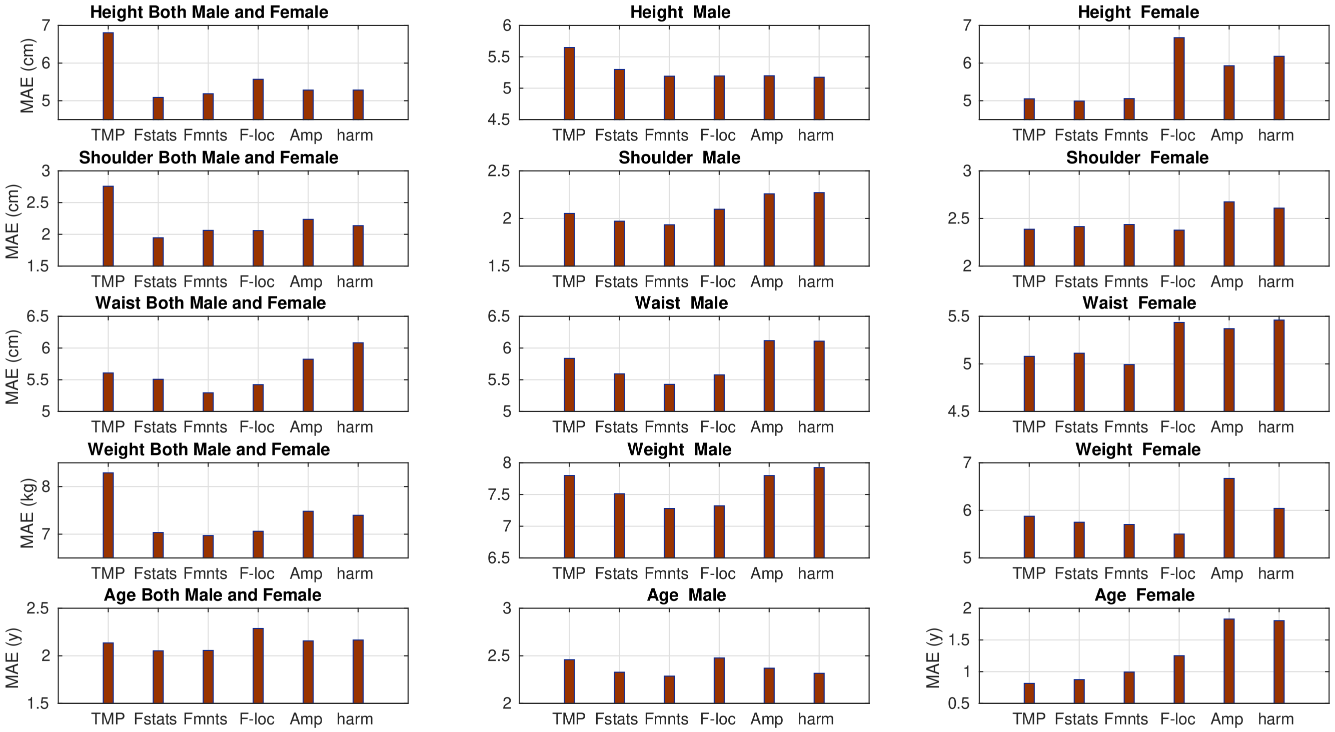

Figure 4.4: Comparison of TMP MAE with multiple physical parameters MAE of different features (i.e, first order statistics(Fstats), formants (Fmnts), harmonic frequency locations (F-loc), amplitude (Amp) and harmonic features (harmonic frequency locations & amplitude features together: harm)), using AFDS.

Table 4.7: Statistics of Train and Test splits of each physical parameter in the AFDS when both gender speakers are considered.

Physical Characteristic	**Min**	**Max**	**Mean**	**Standard Deviation**
		Train Speakers		
Height (*cm*)	147	188	167.4	8.6
Shoulder width (*cm*)	30	53	43.4	3.9
Waist size (*cm*)	64	112	84.9	8.0
Weight (*kg*)	39.4	106.9	64.4	12.8
Age (*y*)	18	31	23.18	2.91
		Test Speakers		
Height (*cm*)	149	188	169.1	8.1
Shoulder width (*cm*)	36	50	43.7	3.4
Waist size (*cm*)	65	105	84.3	7.4
Weight (*kg*)	46	93.8	64.54	10.7
Age (*y*)	18	37	23.67	3.04

respectively when both gender speakers are considered. Similarly, in case of formants the MAE improvement with TMP when both gender speakers are considered is 23.8%, 25.2%, 5.6%, 15.9% and 3.7% for height, shoulder size, waist size, weight, and age respectively. The harmonic features shows an improvement of 22% better than TMP MAE for height, shoulder size estimation, and 10% improvement in weight estimation when both gender speakers are considered.

In the case of male speakers, Fstats and formants there is an improvement over TMP MAE is around 5% and 8% respectively for all the parameters. The harmonic features are performing better with height and age estimation by 8.4% and 5.8% respectively when compared with TMP MAE. Whereas for female speakers, weight and height estimations shows an improvement around 2% over TMP MAE when Fstats are considered. Formants have better performance of around 3% in waist and weight estimations. Harmonics are not performing better than TMP MAE in any of the physical parameters.

4.5.2 Feature Combination Results

As mentioned in Section 3.8.3, the simple average of these features are performing better with TMP for each physical parameter.

Simple averaging is performed on the predicted test targets obtained from Fstats, formant, and harmonics features. The comparison of combination results with training data mean predictor are listed in Table 4.8. All the results use the same train and test split described in Section 4.5. The performance metrics both MAE and RMSE on Comb -3 are better than the TMP when both gender speakers and male speakers are considered except in waist estimation. In the waist estimation Comb -2 is performing better than TMP. In case of female physical parameter estimation, each parameter shows better performance with different combinations of features. Comb -3 performs better with weight and age estimation, whereas Comb -1 and Comb-2 perform better with shoulder and waist estimation, respectively. Any of the feature combinations didn't perform well in the height estimation of a female speaker.

We hypothesize that female speakers' training data (33 speakers) is insufficient to train the model. To verify this, we have added the training data of TIMIT female speakers along with AFDS female training data to the train the SVR model and tested on the AFDS female data to estimate the speaker's height. This showed an improvement in the MAE from 5.4 cm to 4.6 cm for female speakers. However, we cannot do the same for other physical parameters as the TIMIT dataset has only height and age information.

4.5.3 Duration Analysis

To analyze the minimum amount of speech required for estimating the physical parameter in a multilingual setting, we evaluate the system's performance at different utterance durations.

We extend the same duration analysis (please refer Section 3.8.4) on all the physical parameters[5] of AFDS. The system performance is evaluated for different lengths of speech files ranging from 0.25s to full duration(around 40s). We observed that mean absolute errors of each physical parameter for different durations' of the speech signal is less than the TMP for both gender speakers and male speakers except in shoulder size estimation. In male speakers' shoulder estimation by using 0.5s speech data, the prediction error (MAE) is less than the MAE of TMP. From this, it is evident that the

[5] Height, Age, Shoulder width, Waist size, and Weight

Table 4.8: Comparison of the proposed feature combinations – Comb -1 (Fstats + formant + frequency locations), Comb -2 (Fstats + formant + amplitude), Comb -3 (Fstats + formant + harmonic features (amplitude + frequency locations)) with TMP of AFDS.

Multiple Physical parameter Estimation – All (Male + Female)

	TMP		Comb-1		Comb-2		Comb-3	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Height(cm)	6.8	8.2	5.1	6.3	5.0	6.1	**5.0**	**6.1**
Shoulder(cm)	2.8	3.4	2.0	2.4	2.0	2.4	**1.9**	**2.4**
Waist(cm)	5.6	7.3	**5.3**	**6.9**	5.4	6.9	5.5	7.0
Weight(kg)	8.3	10.6	6.9	9.0	7.0	8.9	**6.9**	**8.8**
Age (y)	2.1	3.1	2.1	3.0	2.0	2.9	**2.0**	**2.9**

Multiple Physical parameter Estimation – Male

	TMP		Comb -1		Comb -2		Comb-3	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Height(cm)	6.4	6.9	5.1	6.3	5.1	6.2	**5.0**	**6.1**
Shoulder(cm)	2.1	2.5	2.0	2.4	2.0	2.4	**2.0**	**2.4**
Waist(cm)	5.8	7.3	**5.4**	**7.0**	5.6	7.1	5.5	7.1
Weight(kg)	7.8	9.6	7.3	9.2	7.4	9.2	**7.4**	**9.1**
Age (y)	2.5	3.4	2.3	3.3	2.2	3.1	**2.2**	**3.1**

Multiple Physical parameter Estimation – Female

	TMP		Comb -1		Comb -2		Comb-3	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Height(cm)	**5.1**	**5.9**	5.6	6.3	5.3	6.2	5.4	6.3
Shoulder(cm)	2.4	2.9	**2.4**	**2.9**	2.5	3.1	2.4	3.0
Waist(cm)	5.1	7.2	5.0	7.1	4.9	**6.5**	**4.9**	6.6
Weight(kg)	5.9	8.4	5.6	8.2	5.7	8.3	**5.5**	**8.1**
Age(y)	0.8	1.0	0.9	1.1	1.0	1.2	**0.7**	**0.9**

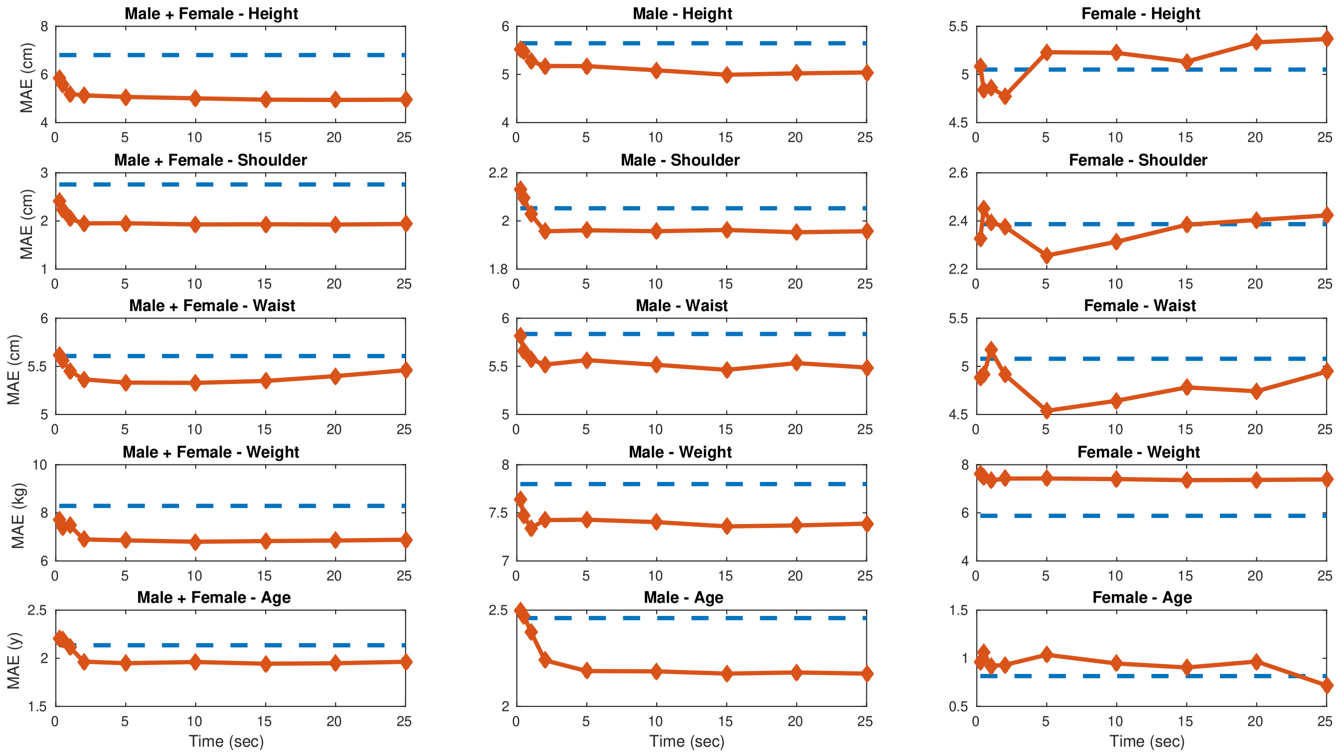

Figure 4.5: MAE vs duration of utterance, for physical parameters' (Height, Shoulder width, Waist size, Weight and Age) estimation from AFDS database. The horizontal dashed line represent training data mean predictor (TMP) benchmark.

system is reliably able to predict the physical parameters from 0.5s duration of the speech signal with prediction error less than the training data mean. The duration of speech at which the prediction error saturates is around 2s when both genders' data is considered together. When there is 2s of speech data for both gender speakers, the mean absolute error for height is 5.1cm, shoulder width is 1.9cm, waist size is 5.4cm, weight is 6.9 kg and for age is 2 years. Whereas when the available speech data is 40s, we have 5.0 cm, 1.9cm, 5.5cm, 6.9kg and 2y for height, shoulder width, waist size, weight , and age respectively when both gender speakers are considered. The variation of MAE with respect to utterance duration for both genders, male and female speakers are shown in Figure.4.5. For male speakers, the MAE saturates around 2s, like the above mentioned case (both genders). The change in MAE when full duration (40s) and 2s considered is 0.1cm in height, and there is no change in MAE for other physical parameters like shoulder size, waist size, weight and age estimation. As mentioned above in previous section, as the number of training female speakers is less, the prediction error in multilingual setting using Comb-3 set of features does not help much in duration analysis.

4.5.4 Effect of Language

To understand language's effect, we perform the physical parameter estimation in the multilingual setting by splitting the English utterances and native language utterances into train and test splits. There are 489 English utterances out of which 119 female speaker utterances and 370 male speakers utterances for training split. There are 272 English utterances for test split, in that 52 are female utterances, and 220 are male utterances. Similarly, 462 native language utterances in that 105 are female speaker utterances and 357 male speaker utterances for train split. There are 266 native language utterances for test split, out of which 52 are female utterances, and 214 are male native language utterances.

The Comb–3 scheme has the least error in speaker profiling. Thus this scheme is used to evaluate the robustness of the system to the language. Hence, the system is trained and evaluated with two different subsets of the data

1. Native language utterances.

2. English utterances.

The support vector regression model is trained using one of the subsets at a time for each physical parameter estimation. The system is then evaluated using the matched as well as mismatched subset.

1. **Matched Condition:**
   Case-1: The SVR models are trained using the English utterances only, and is evaluated separately on the English utterances.
   Case-2: The SVR models are trained using the native languages and the system is evaluated with the native languages utterances only.

2. **Mismatched Condition:**
   Case-1: The SVR models are trained using the English utterances only, and is evaluated separately on the native languages utterances.
   Case-2: The SVR models are trained using the native languages and the system is evaluated with the English utterances only.

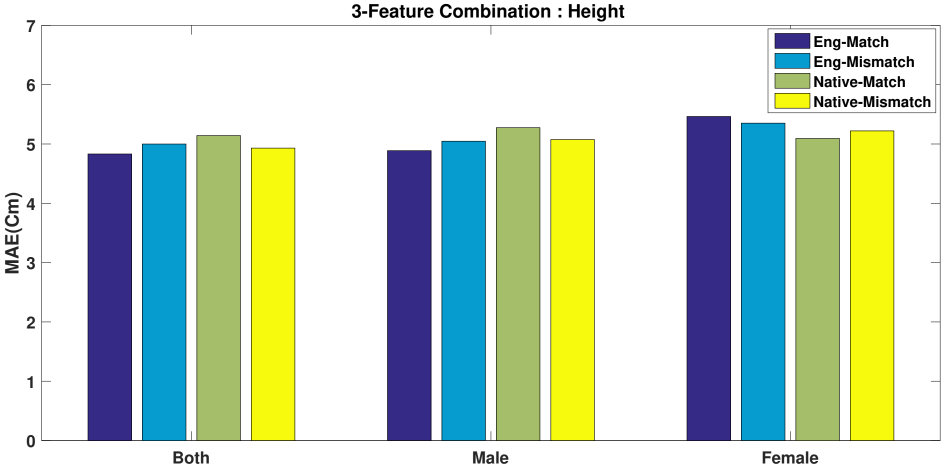

Figure 4.6: Gender wise MAE comparison of matched and mismatched conditions in height estimation using Comb-3 set of features on AFDS

We reported the system performance using Comb–3 features set for the matched and mismatched cases. We presented the system degradation from the perspective of mismatched cases for each physical parameter.

Height Estimation: When the system is evaluated on mismatched condition case-1, the maximum system performance degradation is of 3% in male speakers with the Comb-3 set of features in height prediction. In contrast, there is no performance degradation with the Comb-3 set of features for female and both gender speakers.

Similarly, when a model is evaluated on mismatched condition case-2, the system can estimate the height of male speakers as well as both gender speakers without degrading the system performance by using Comb-3 set of features. In the case of female speakers, Comb-3 set of features deteriorates the system performance by 2.5%. The gender-wise height MAE of matched and mismatched conditions for Comb-3 set of features are shown in Figure 4.6.

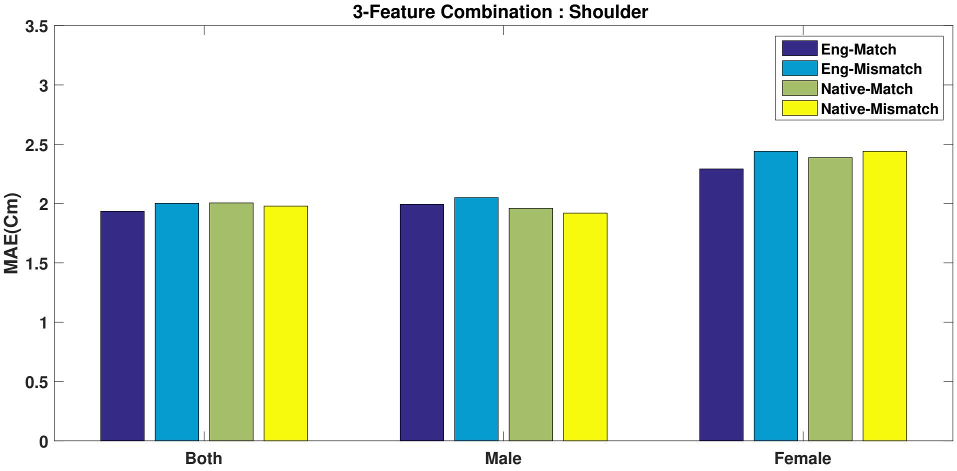

Figure 4.7: Gender wise MAE comparison of matched and mismatched conditions in shoulder size estimation using Comb-3 set of features on AFDS

Shoulder size Estimation: When the system is evaluated on mismatched condition case-1, the maximum system performance degradation when both gender speakers are considered is 3.5% with Comb-3 set of features. In male speakers, the Comb-3 set of features degrades the shoulder estimation system by 3% and 6.5% for female speakers.

Similarly, when the system is evaluated on mismatched condition case-2 while predicting a speaker's shoulder size, the system can predict for both gender speakers as well as male speakers without degrading the performance by Comb-3 set of features. Whereas, when female speakers are considered Comb-3 set of features degrade the system by 2.2%. The gender-wise shoulder size MAE of matched and mismatched conditions for Comb-3 set of features are shown in Figure 4.7.

Waist size Estimation: When the system is evaluated in mismatched condition

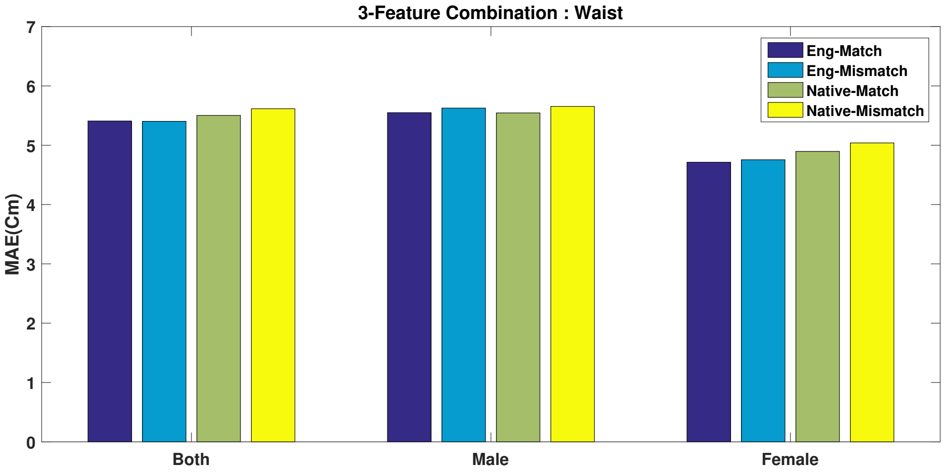

Figure 4.8: Gender wise MAE comparison of matched and mismatched conditions in waist size estimation using Comb-3 set of features on AFDS

case-1, while predicting the male and female speaker's waist, the maximum system performance degradation is 1% with Comb-3 set of features. In contrast, there is no degradation in system performance when both gender speakers are considered.

Similarly, when the system is evaluated in mismatched condition case-2, the system shows the degradation of 3% with the Comb-3 set of features for female speakers. At the same time, it is 2% for male speakers and both gender speakers. The gender-wise waist size MAE of matched and mismatched conditions for Comb-3 set of features are shown in Figure 4.8.

Weight Estimation: When the system is evaluated in mismatched condition case-1, while predicting a speaker's weight, the system degrades the system performance by 1% for both genders and female speakers with Comb-3 set of features. At the same time, it is 2% for male speakers with the Comb-3 set of features.

Similarly, when the system is evaluated in mismatched condition case-2, the system shows the degradation of maximum of 1% with Comb-3 set features for all the speakers. The gender-wise weight MAE of matched and mismatched conditions for Comb-3 set of features are shown in Figure 4.9.

Age Estimation: When the system is evaluated in mismatched condition case-1, the maximum performance degradation while predicting the age is 3% and 4% for both genders and male speakers, respectively, with the Comb-3 set of features. At the same time, there is no performance degradation in female speakers.

Similarly, when the system is evaluated in mismatched condition case-2, the system

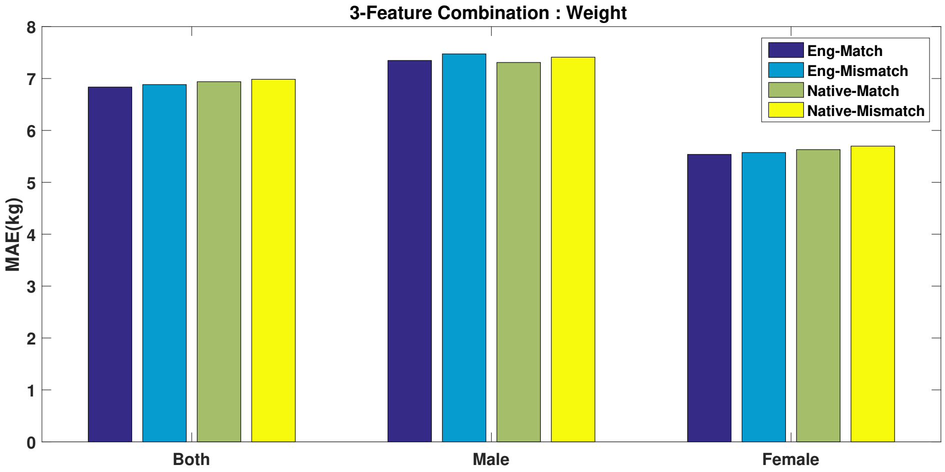

Figure 4.9: Gender wise MAE comparison of matched and mismatched conditions in weight estimation using Comb-3 set of features on AFDS

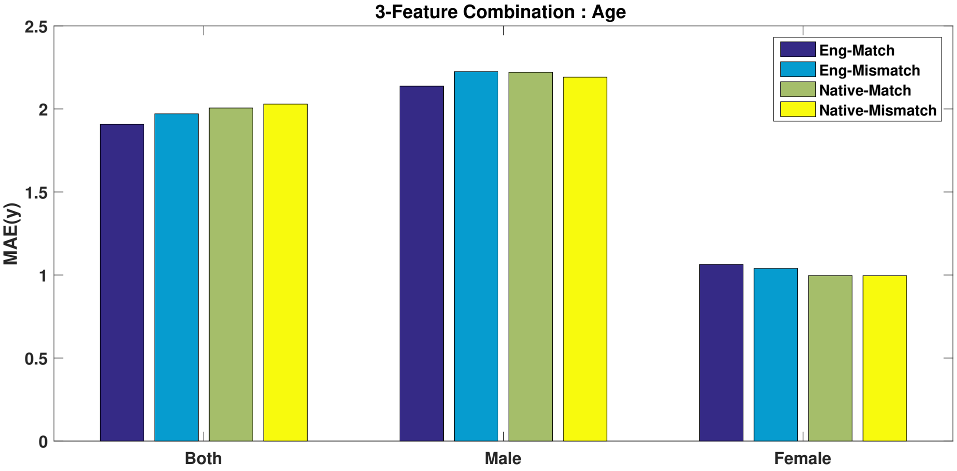

Figure 4.10: Gender wise MAE comparison of matched and mismatched conditions in age estimation using Comb-3 set of features on AFDS

shows the degradation of 1.2% with Comb-3 set of features when both gender speakers are considered. At the same time, the Comb-3 set of features can predict the age of male and female speakers with the same trained model without degrading the performance. The gender-wise age MAE of matched and mismatched conditions for Comb-3 set of features are shown in Figure 4.10.

In summary, the maximum degradation for male speakers in case of age estimation is 4%, for height and shoulder it is 3%, and for waist and weight estimation it is 2%. The maximum performance degradation for female speakers in case of shoulder estimation is 6.5%, height is 2.5%, waist size is 3% and for weight estimation it is 1%. Similarly when both gender speakers are considered, the performance degradation is of 3% for age, 3.5% shoulder size, 2% for waist size and 1% for weight estimation, and no degradation in height estimation.

4.6 Experimental Results on NISP dataset

Multiple physical parameters are estimated in a multilingual setting on the NISP dataset using the same approach to estimate height and age using the TIMIT dataset. For evaluation purposes, the dataset is divided into train and test splits without overlapping any speakers. The training split has 210 speakers in which 134 male and 76 female speakers. The train split has 17161 utterances, out of which 10911 utterances are from male speakers, and 6250 utterances are from female speakers. The test split has 135 speakers out of which 85 speakers are male and 50 are female speakers. This test split has 11107 utterances, which includes 6933 male speaker utterances and 4174 female speaker utterances. The statistics of train and test splits of the dataset are given in Table 4.9. As mentioned before, the standard error metrics mean absolute error and root mean square error are used to measure the errors from the actual and predicted targets.

Table 4.9: Statistics of Train and Test splits of each physical parameter in the NISP dataset when both genders are considered.

Physical Characteristic	**Min**	**Max**	**Mean**	**Standard Deviation**
		Train Speakers		
Height (*cm*)	143	191	167.1	9.5
Shoulder width (*cm*)	32	55	42.9	4.2
Weight (*kg*)	36.9	116.5	65.4	14.0
Age (*y*)	18	47.5	24.8	6.0
		Test Speakers		
Height (*cm*)	146.5	182.5	166.7	8.5
Shoulder width (*cm*)	30.0	53.0	42.9	3.7
Weight (*kg*)	34.1	93.8	63.5	11.3
Age (*y*)	18.3	43.6	24.4	5.5

4.6.1 Individual Feature Results

We evaluate the system using the each of the features separately. The MAE of each feature is shown in Fig 4.11. This is compared with the default approach – target mean predictor (predicting the target of each physical parameter using the mean of training data of each parameter).

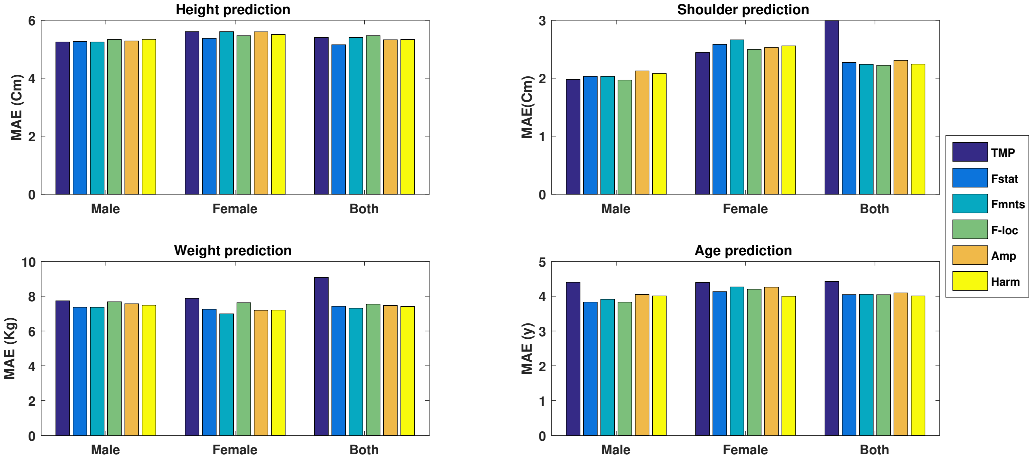

Figure 4.11: Gender wise MAE of each feature (Fstat, Formants (Fmnts), frequency locations (F-loc), Amplitude (Amp) and Harmonic features (amplitude + frequency locations – Harm)) compared with Training data Mean Predictor (TMP) of the NISP dataset

The below figure shows a clear improvement in the MAE in all the physical parameters for both gender speakers except in height when harmonics frequency locations are considered. In height estimation, the maximum improvement in MAE is 4.7%, 1.8%, and 4.2% over TMP when both gender speakers, male and female speakers, are considered respectively. In the case of shoulder estimation, all features show a minimum improvement of 22.1% over TMP in MAE when both gender speakers are considered but not with male and female speakers. In weight estimation, MAE improvement is 4.7%, 11.3%, and 19.5% over TMP when male, female, and both gender speakers are considered, respectively. In the case of age estimation, all the features showed an improvement of MAE over TMP. The improvement in MAE is 13%, 9%, and 10.2% for male, female, and both gender speakers are considered in estimating a speaker's age.

However, the degradation of MAE over TMP in estimating male speakers height and shoulder is maximum of 2% and 7.4%, respectively. In female speakers, none of the features showed improvement over TMP and there is a degradation of MAE over TMP is of 2% minimum and a maximum of 9% in estimating shoulder size of a speaker.

4.6.2 Feature Combination Results

We combined the predicted targets from three different Support Vector Regression outputs to improve the final physical parameter estimates. We have made three different sets of feature combinations of Fstats and formant features with either harmonic frequency location (Comb–1) or amplitude (Comb–2) or harmonic features (both frequency and amplitude features Comb–3). These results are tabulated in comparison with default predictor (TMP) in Table 4.10. This simple average of these features' regressed predicted targets has improved the predicted error metrics over the individual error metrics. The MAE and RMSE of both gender speakers improved relatively by about $22 - 29\%$ in body build parameter estimation (height, shoulder width and weight) tasks using the Comb–3 set of features. Similarly, in age estimation, we observe a relative improvement of 14% improvement in MAE. There is a relative improvement over the TMP with three feature combination (comb–3) in all the physical parameters except in RMSE of female speakers' shoulder size and male speakers' age.

4.6.3 Duration Analysis

To analyze the minimum amount of speech required for estimating the physical parameter in a multilingual setting on the collected NISP dataset. The same train and test splits of the NISP dataset are also considered for the duration analysis (please refer Section 4.6).

We evaluate the performance of the system at different utterance durations. Comb-3 set features are used to perform the duration analysis on collected physical parameters[6] using the NISP dataset. The system performance is evaluated for different lengths of speech files ranging from 0.25s to full duration(around 40s) with an average duration of 10s.

We observed that each physical parameter's mean absolute errors for different durations' of speech utterance are less than the TMP MAE for all the speakers except

[6] Height, age, shoulder width, and weight of a speaker

Table 4.10: Comparison of the proposed feature combinations – Comb -1 (Fstats + formant + frequency locations), Comb -2 (Fstats + formant + amplitude), Comb -3 (Fstats + formant + harmonic features (amplitude + frequency locations)) with TMP of NISP dataset.

Height (cm) Estimation						
	Male		Female		All	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
TMP	5.22	6.17	**5.30**	6.93	7.14	8.47
Comb–1	5.20	**6.07**	5.42	6.77	5.21	6.23
Comb–2	**5.12**	6.10	5.35	6.74	5.07	**6.13**
Comb–3	5.16	6.14	5.32	**6.70**	**5.12**	6.17
Shoulder (cm) Estimation						
TMP	1.98	2.58	**2.44**	**3.52**	2.99	3.73
Comb–1	1.95	**2.50**	2.53	3.64	2.16	2.90
Comb–2	1.96	2.51	2.48	3.56	2.15	2.89
Comb–3	**1.95**	2.51	2.50	3.59	**2.14**	**2.89**
Weight(kg) Estimation						
TMP	7.74	9.57	7.88	9.76	9.08	11.35
Comb–1	7.33	9.03	7.15	8.98	7.27	8.99
Comb–2	7.11	8.80	6.89	8.70	7.13	8.85
Comb–3	**7.10**	**8.84**	**6.89**	**8.68**	**7.11**	**8.85**
Age(y) Estimation						
TMP	4.40	**5.60**	4.39	**5.57**	4.42	**5.54**
Comb–1	3.79	5.67	4.09	6.05	3.96	5.86
Comb–2	3.80	5.64	4.06	5.99	3.92	5.81
Comb–3	**3.80**	5.65	**3.98**	5.95	**3.90**	5.80

in female speakers' height and shoulder size. The Figure 4.12 shows the MAE performance of each physical parameter versus the target mean predictor. From the plot it is very clear that the physical parameter can be estimated with a minimum of 0.25s of speech data expect height and shoulder size of female speakers. The prediction error gets saturated at 2s of speech data when both gender speakers' speech data is considered. At the same time, it is saturating at 5s of speech data in case of male speakers all physical parameters and age and weight estimation in female speakers.

We hypothesis that the number of speakers for female speakers is less, and the variability in actual targets is also less. Most of the predicted targets are skewed towards the mean of the physical parameter of female speakers.

4.6.4 Effect of Language

In order to understand the effect of language in the physical parameter estimation system using NISP dataset, the SVR model is trained separately with native language utterances and English utterances. The physical parameter estimation system is trained and tested in matched and mismatched conditions of the training and testing utterances. The details of matched and mismatched conditions are detailed in Section 4.5.4.

There are 8245 native language utterances in the train split, out of which 5236 are male speaker utterances, and 3009 are female speaker utterances. There are 5775 native utterances for test split, out of which 3587 are male and 2188 female speaker utterances.

Similarly, 8916 English utterances for train split, out of which 5675 are male, and 3241 are female speaker utterances. For test split, 5332 English utterances, out of which 3346 are male utterances, and 1986 are female utterances.

Height Estimation: When the system is trained on English utterances and tested on native language utterances (mismatched condition case-1) in predicting a speaker's height, the maximum system performance degradation is about 3% with the Comb-3 set of features for all the speakers.

Similarly, when the system is trained on native language utterances and tested on English utterances (mismatched condition case-2), the Comb-3 set of features do not degrade the system's performance for all the speakers. The system can predict a speaker's height reliably from English utterances, even though it is trained on multiple languages for male, female, and both gender speakers. The gender-wise height

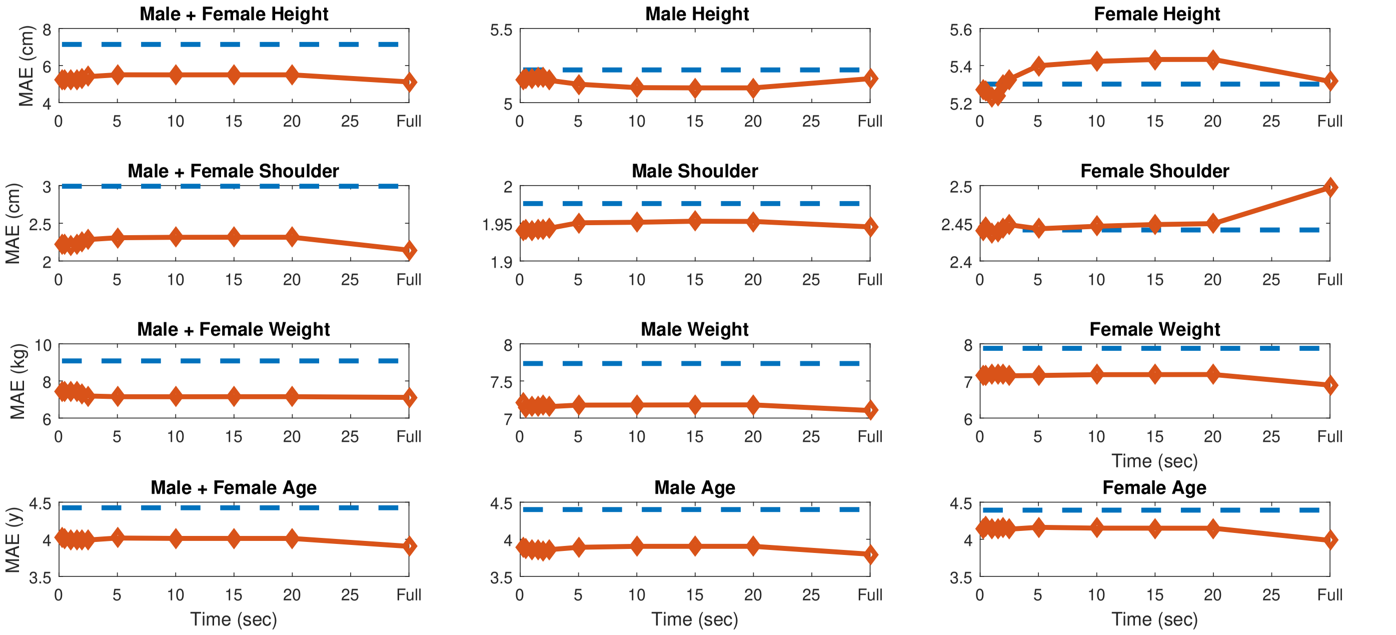

Figure 4.12: MAE of Comb–3 vs duration of utterance, for physical parameters' (Height, Shoulder width, Weight, and Age) estimation from NISP database. The horizontal dashed line represent training data mean predictor (TMP) benchmark.

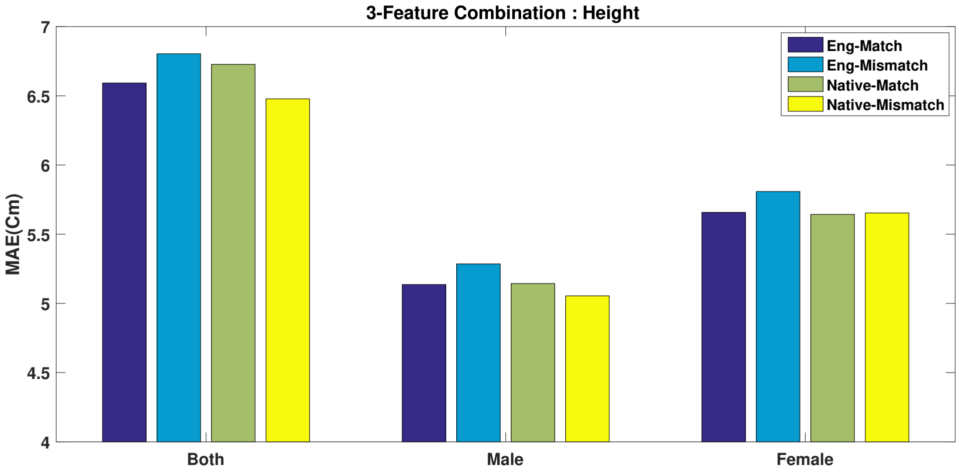

Figure 4.13: Gender wise MAE comparison of matched and mismatched conditions in height estimation using Comb-3 set of features on NISP dataset

MAE of matched and mismatched conditions for Comb-3 set of features are shown in Figure 4.13.

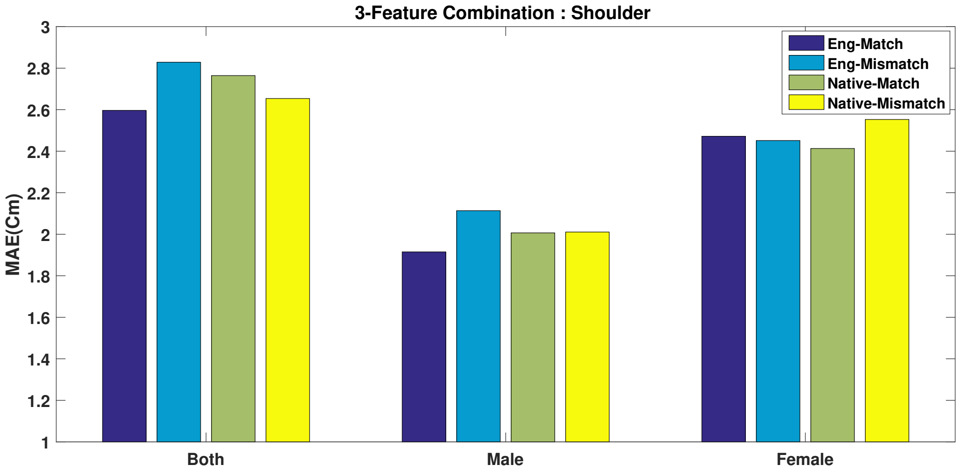

Figure 4.14: Gender wise MAE comparison of matched and mismatched conditions in shoulder estimation using Comb-3 set of features on NISP dataset

Shoulder size Estimation: When the system is evaluated for mismatched condition case-1, while predicting a speaker's shoulder size, the maximum system performance degradation is about 10% with Comb-3 set of features when male speakers utterances are considered. In the case of both gender speakers, Comb-3 set of features degrade the system performance by 9%. The trained model can predict the female speaker's shoulder size without degrading the system's performance.

Similarly, when it is evaluated on mismatched condition case-2, the system degrades the performance by 6% for female speakers when the Comb-3 set of features are considered. There is no performance degradation for male speakers as well as both gender speakers with Comb-3 set of features. The gender-wise shoulder size MAE of matched and mismatched conditions for Comb-3 set of features are shown in Figure 4.14.

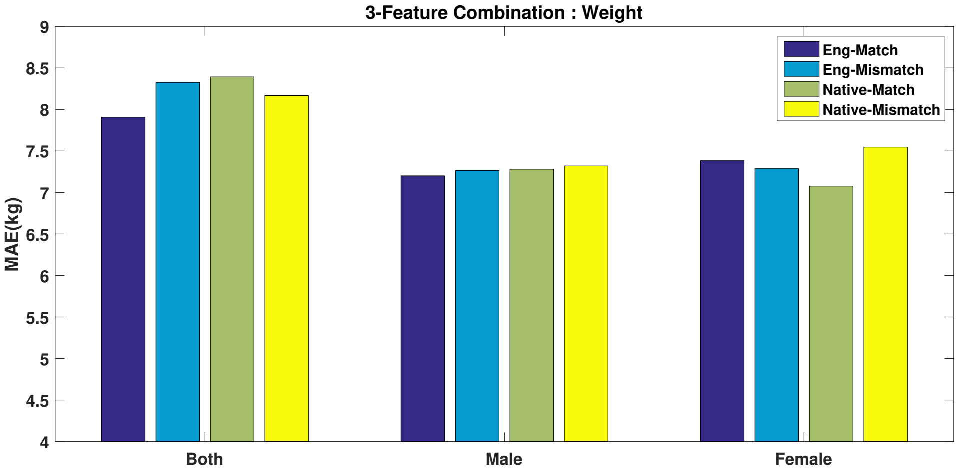

Figure 4.15: Gender wise MAE comparison of matched and mismatched conditions in weight estimation using Comb-3 set of features on NISP dataset

Weight Estimation When the system is evaluated on mismatched condition case-1, while predicting a speaker's weight, Comb-3 set of features degrades the system by 1% for males and 6% when both gender speakers are considered. In female speakers, without degrading the system performance, the weight of a speaker can be predicted.

Similarly, when a model trained on native language utterances and tested on English utterances (mismatched condition case-2), the performance degradation is about 6.6% and 0.5% with Comb-3 set of features when female and male speakers are considered respectively. Whereas, there is no degradation when both gender speakers are considered. The gender-wise weight MAE of matched and mismatched conditions for the Comb-3 set of features are shown in Figure 4.15.

Age Estimation: When the system is evaluated on mismatched condition case-1, while predicting the age of a speaker, the system degrades male speakers' performance

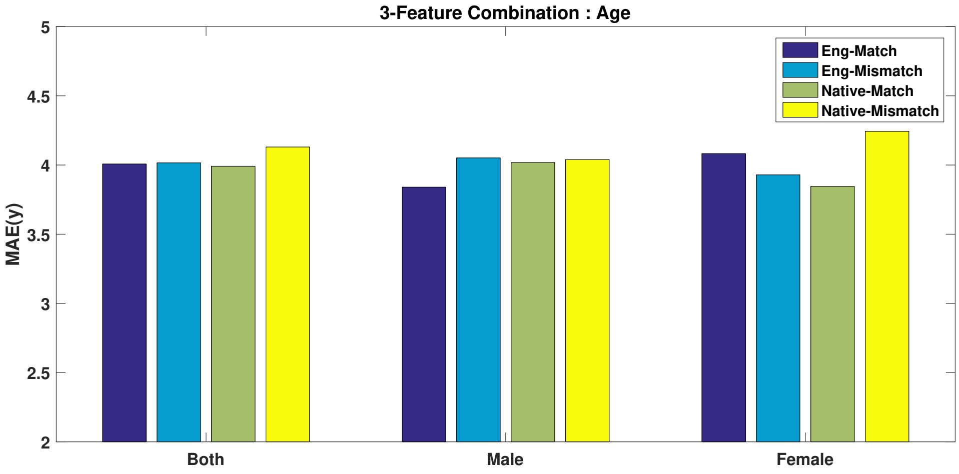

Figure 4.16: Gender wise MAE comparison of matched and mismatched conditions in age estimation using Comb-3 set of features on NISP dataset

by 5.5% when Comb-3 set of features is considered. In the case of both gender speakers and female speakers, Comb-3 set of features can predict a female speaker's age without degrading the system performance.

Similarly, when a model trained on native language utterances and tested on English utterances (mismatched condition case-2), Comb-3 set of features degrades the system performance by 10% for females, 0.5% for males, and 3.5% when both gender speakers are considered. The gender-wise weight MAE of matched and mismatched conditions for Comb-3 set of features are shown in Figure 4.16.

In short, the maximum degradation in the system performance in the mismatched conditions of male speakers in height estimation is 3%, shoulder is 10%, weight is 1% and age of a speaker is 5.5%. In female speakers, the degradation of the performance is 3% for height, 6% for shoulder, 6.6% for weight and 10% for age estimation. Similarly, in the case of both genders, the degradation in the system performance is 3% in height, 9% in shoulder size, 5% in weight and 3.5% in age estimation.

4.7 Summary

As a summary of this chapter, we have addressed the physical parameter estimation system in a multilingual setting. Two different datasets, AFDS and NISP multilingual and multi-accent datasets have been collected for the speaker's physical parameter estimation. These datasets have 207 and 345 distinct speakers, respectively. AFDS has 1489 utterances, each of 40s length, whereas the NISP dataset has 28268 utterances with an average length of 10s. AFDS has 8.4 hours of English speech data and 7.91 hours of native language speech data; overall, AFDS has 16.3 hours of speech data. In the case of NISP dataset, there are 24.83 hours of native language speech data and 32.03 hours of English speech data, as the overall NISP dataset has 56.86 hours of speech data. AFDS has physical parameters details like height, age, shoulder size, waist size, and weight of a speaker are estimated from AFDS, whereas the NISP dataset has height, age, shoulder size, and weight of a speaker.

Different sets of multi-resolution features have explored (such as Mel filter bank features, formants, frequency locations, amplitude, and harmonic features) in estimating a speaker's multiple physical parameters in the multilingual setting.

Comb–3 set of features shows a significant improvement in physical parameter MAE over TMP MAE on AFDS when both gender speakers are considered by 26% in height, 32% in shoulder size, 17% in weight, and 5% in age estimation of a speaker. In the case of male speakers, there is an improvement of 5% MAE compared with TMP MAE for shoulder size, waist size and weight of a speaker, whereas it is 22% in height estimation and 12 % in age estimation. Comb–3 set of features showed an improvement in weight and age estimations for female speakers.

The duration analysis was performed using Comb–3 set of features for estimating the physical parameter in a multilingual setting. The prediction error gets saturated at 2s of speech data for male speakers, and both gender speakers are considered.

The effect of language is also studied on AFDS by training the model with matched and mismatched conditions of speech utterances. The system degrades utmost by 4% for males, 6.5% for females, and 3.5% when both gender speakers are considered across all physical parameters with Comb-3 set of features in mismatched conditions.

Similarly, Comb–3 set of features shows a significant improvement over TMP in NISP dataset too, when both genders are considered in height, shoulder size, weight, and age estimations. There is an improvement of 28% in height, 21% in shoulder size, and 12% in weight and age estimation of a speaker over TMP MAE when both gender speakers are considered.

The duration analysis is also performed in NISP dataset using Comb–3 set of features in a multilingual setting. It is observed that with a minimum of 0.25s of speech data, all the physical parameters can be predicted except the female speaker's height and shoulder size. The prediction error gets saturated at 2s of speech data when both gender speakers speech utterances are considered and it is at 5s for male speakers. Similarly, in understanding the effect of language, when the model is trained either with only one language (English) or with multiple languages (5 languages), the utmost degradation of the prediction is about 10% for all the physical parameters when both gender speakers are considered.

The collected datasets have potential speaker profiling, accent- language identification, speaker recognition, and speech recognition.

Chapter 5

End to End Physical Parameter Estimation System

This chapter aims to jointly predict all the physical parameters of a speaker using a single system from short-duration ($1 - 3$s) speech inputs. We propose a DNN architecture to jointly predict speaker parameters like age,height, shoulder size, waist size, and weight from speech data. We explore a novel scheme to initialize the network using a conventional system based on support vector regression trained with GMM-UBM super-vector features. This initialization eliminates the need for large amounts of data for the deep neural network training. To the best of our knowledge, this is the first attempt to develop an end-to-end model that predicts the multiple parameters of a speaker jointly. We evaluate the system on collected multilingual and multi-accent AFDS, and NISP datasets for predicting the age and body build parameters like height, shoulder size, waist size, and weight of a speaker.

The highlights of this chapter can be summarized as follows:

1. We propose a unified DNN architecture to predict both age and body build parameters like height, shoulder size, waist size and weight of a speaker for short durations of speech.

2. A novel initialization scheme for the deep neural architecture is introduced, that avoids the requirement for a large training dataset.

3. We evaluate the system in predicting the height and age of a speaker on standard TIMIT dataset where the mean duration of speech segments is around 2.5s.

4. We also evaluate the system on collected multi-lingual and multi-accent AFDS and NISP datasets in predicting the age as well as body build parameters like height, shoulder size, waist size and weight of a speaker.

The rest of this chapter is organized as follows. The Section 5.1, details about the baseline system, statistical representation of features and support vector regression model to estimate the physical parameter of a speaker. Section 5.2 details the proposed deep neural network architecture for physical parameter estimation. The experimental results on TIMIT dataset in predicting height and age of a speaker are detailed in Section 5.3 along with initialization schemes and error analysis. Extending the joint prediction of multiple physical parameter estimations on the collected multilingual and multi accent datasets are detailed in Section 5.4. Finally, Section 5.5 reports the key findings and summary of the chapter.

5.1 Baseline system

We use the system from Chapter 3 Section 3.3.2 as the baseline. Our baseline system is trained with linear support vector regression model using first order statistics computed from a GMM model. We train a GMM-UBM with diagonal covariance using cepstral features of the train data. For a given the sequence of input feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$, the density function of GMM is given by,

$$p(\mathbf{x}) = \sum_{k=1}^{M} w_k \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_k, \mathbf{C}_k), \tag{5.1}$$

where $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$, $\boldsymbol{\mu}_k$ denote the input feature vector and mean respectively and $\boldsymbol{C}_k$ represents diagonal covariance matrix of the k^{th} GMM component with weight w_k. The frame level first order statistics (defined as $\mathbf{f}_i^j$ for a given frame i is computed as,

$$\mathbf{f}_i^j = \mathbf{x}_i p(j|\mathbf{x}_i), \tag{5.2}$$

where the *a-posterior* probabilities $p(j|\mathbf{x}_i)$ are computed by the Bayesian rule, given as follows,

$$p(j|\mathbf{x}_i) = \frac{w_j \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}_j, \mathbf{C}_j)}{\sum_{k=1}^{M} w_k \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}_k, \mathbf{C}_k)}. \tag{5.3}$$

We concatenate all mixture component specific stats $\mathbf{f}_i^j$ to form a frame level super vector $\mathbf{F}_i$. We perform the mean across time to get first order statistics $\mathbf{F}$ (referred as Fstats) across the entire speech utterance.

$$\mathbf{F} = \frac{1}{T} \sum_{i=1}^{T} \mathbf{F}_i \tag{5.4}$$

The vector $\mathbf{F}$ (called Fstats) is used as the feature representation for the regression system (Babu and Vijayasenan (2017)). Separate SVR models are trained to predict the physical parameters. As the dimension of the input features is high, we used a linear SVR. The prediction model of the linear SVR for an input frame $\mathbf{x}_i$ is given by,

$$H_i = \sum_{i=1}^{n_s} \mathbf{v}_i^T \mathbf{F} + b = \mathbf{w}^T \mathbf{F} + b \tag{5.5}$$

where n_s is the number of support vectors $\mathbf{v}$, b is the bias, and $w = \sum_{i=1}^{n_s} \mathbf{v}_i$. The prediction output H_i indicates the physical parameter estimate for the current feature vector $\mathbf{x}_i$. The average prediction (averaged over the frames $1 \ldots T$) is used as the estimate of physical parameter for the utterance. The SVR models are trained and evaluated separately for male and female speakers.

5.2 Deep Neural Network Architecture for Joint Prediction of Physical Parameters

The proposed deep neural architecture for joint prediction of age and body build parameters (height, shoulder size, waist size and weight) of a speaker is inspired from our baseline algorithm. The block diagram of the proposed DNN model is shown in Figure. 5.1.

The model has three parts. The first part (Layers L1, L2, L3) corresponds to GMM posterior computation (Eq: 5.3). The second part (Layers L4, L5, L6) performs statistics computation (Eq: 5.4) and the final part (Layer L7) represents the SVR regression (Eq: 5.5). The first part is a fully connected multilayer perceptron shared among all the input speech frames. The second part performs frame wise first order statistics computation and computes the mean along time to get the statistics across the entire speech utterance. The trainable parameters of the network are in Part 1, 3.

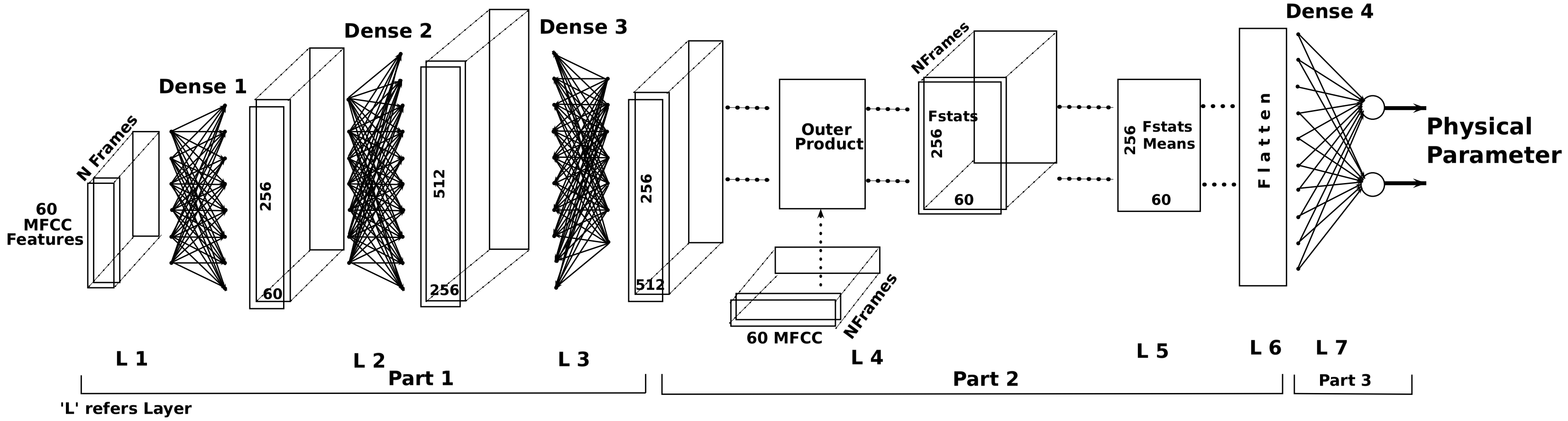

Figure 5.1: Block diagram of deep neural network architecture for joint prediction of physical parameters (age and body build parameters (height, shoulder size, waist size and weight)) of a speaker from speech

Typically, deep neural network (DNN) architectures require a lot of training data to learn the parameters. Further, the model has to be efficient to perform regression on very short duration variable length speech segments. We exploited our baseline system for an innovative approach for the initialization of the neural network.

Since we envisage the first part of the network to predict the GMM posteriors, we initialize these layers from a smaller network trained to predict the GMM posteriors of the baseline system. A three layer fully connected network is trained separately for this purpose. The network targets for training are obtained as the GMM-UBM frame level posteriors. The network has ReLU non linearities in the hidden layers and softmax at the output layer. The network parameters are learned over the entire training data. The second part of the network exactly replicates the operations performed in Eq. 5.2 and Eq. 5.4 where the posteriors $p(j|\mathbf{x}_i)$ are obtained using the neural network (first part of the network). The third part of the network is about predicting the speaker parameters from the first order statistics. The network is trained with sum of mean square error in age and age and body build parameters (height, shoulder size, waist size and weight) prediction. We initialize this layer from the baseline linear SVR. The weights corresponds to age, height, weight, shoulder size, and waist size of speaker's targets are initialized from the respective SVR models. Following the initialization, the network is trained using back propagation with a mean square error loss. We learn separate models for male and female speakers.

5.3 Experiments and Results on TIMIT Dataset

We perform our experiments on the TIMIT dataset. The standard train-test split is used in all experiments. We consider male and female speakers separately. The details of the dataset are explained in Chapter 3, Section 3.1. There is no overlapping of recordings of speakers in train and test splits. The duration of the recordings ranges from $1-6$s with an average of about 2.5s.

5.3.1 Baseline GMM-UBM-SVR System

We extract 20 Mel Frequency Cepstral Coefficients (MFCC) along with the delta and double delta features (feature dimension 60) from windowed speech. We perform a voice activity detection and cepstral mean and variance normalization for the input

Table 5.1: RMSE values of baseline height and age estimation algorithms

Physical parameter	Singh *et al.* (2016*b*)		Default predictor	
	MALE	FEMALE	MALE	FEMALE
Height(cm)	6.70	6.10	7.01	6.51
Age(y)	7.80	8.90	8.07	9.15
Physical parameter	GMM-UBM-SVR		DNN-postr-SVR	
	MALE	FEMALE	MALE	FEMALE
Height(cm)	6.93	6.30	6.93	6.29
Age(y)	8.22	9.50	8.23	9.50

MFCC coefficients. A 256 component diagonal GMM-UBM is learned from the combined training data of male and female speakers. The first order statistics for each speech utterance is computed as described in Section 5.1. A separate SVR for age and height for male as well as female speakers are learned from Fstats of the training data. We call this method GMM-UBM-SVR. Table 5.1 details the results of this algorithm as well as comparison with a state of the art algorithm on the same task (Singh *et al.* (2016*b*)). The table also lists the results of the default predictor that predicts the training mean value for all test samples. It can be noted that the age prediction algorithm of Singh *et al.* (2016*b*) is only marginally better than the default predictor.

5.3.2 DNN Model Initialization

As detailed in Section 5.2, the first part (Layers L 1 to L 3) is the equivalent of GMM posterior extraction from input MFCC features in the baseline system. Initially, this part is separately trained using the posteriors of GMM-UBM as the target values. The GMM posteriors are computed using Eq. 5.3. The first part network has 2 hidden layers with 256, and 512 hidden neurons and 256 output neurons (corresponding to 256 component GMM). Both hidden layers have a dropout (0.3) and batch normalization operations. The network is trained using back-propagation to minimize the cross-entropy objective function on the TIMIT training data. The training data contains both male and female speakers. This initialization is common for both male and female models.

Table 5.2: RMSE values from DNN model for segment wise and complete duration prediction

Physical parameter	DNN-var-pred		DNN-seg-pred	
	MALE	FEMALE	MALE	FEMALE
Height(cm)	**6.85**	**6.29**	6.87	6.30
Age(y)	**7.60**	**8.63**	7.61	8.65

In order to check the sanity of the trained network, we use the trained DNN posteriors to compute the first order statistics and learn an SVR to predict speaker parameters. We denote the system as DNN-postr-SVR. Table 5.1 presents the corresponding results. It can be seen that the DNN posteriors are attaining very similar performance measures as the GMM-UBM.

The fully connected layer in the third part of the network is initialized from individual linear support vector regression algorithms. The male (female) neural network model is initialized from the male (female) SVR weights for height and age prediction.

5.3.3 DNN Learning

While the network supports variable length inputs for training, we trained it using fixed length speech inputs. We use the Keras toolkit (Chollet *et al.* (2015)) for model learning. We have windowed the input speech into 1 second segments with 0.1-second shift. These short segments along with the corresponding target values are used as the training input for the neural network. The mean square error in height and age is used as the objective function. About 10% of the training data is kept as validation data. The validation performance is used as the training stopping criterion.

The trained network is used for height and age prediction of the test utterances. Note that the test utterances are variable length in nature. This scheme was denoted as DNN-var-pred. Table 5.2 reports the Deep neural network results. It can be seen that the RMSE error of age prediction has improved in both the cases over the DNN-postr-SVR system. The RMSE improvement in case of age prediction is around 0.6 years and 0.9 years for male and female speakers respectively. This is achieved without degrading the RMSE for height prediction.

As a sanity check, we have trained the model without any initialization to the DNN, the error performance is worse than the default predictor (refer Table 5.1). Since the neural network is trained on 1s segments, we also tried to predict the physical parameters using windowed 1-second segments with 0.1-second shift from the variable length speech utterance. The predictions are then averaged to compute the final prediction. The result of this scheme (denoted by DNN-seg-pred) is listed in Table 5.2. The final RMSE values are within ±0.05 of the DNN-var-pred scheme. Thus, even though the network was trained on 1-second length segments, it is able to generalize to variable length speech utterances.

5.3.4 Effect of utterance length

To analyze how shorter segments degrade the performance, we evaluated the GMM-UBM-SVR and DNN-var-pred systems with trimmed speech segments from the test data. We trim the input speech segments to different durations from $1 - 4$ seconds. The variation of RMSE of height prediction with respect to test speech duration is shown in Figure. 5.2. Even with 1 second duration, the degradation in the DNN system performance is 1.7% for male speakers and 3.2% for female speakers. The GMM-UBM-SVR system has an RMSE that is around 0.1cm more than the DNN system for 1s speech input. When the duration increases, the DNN system RMSE

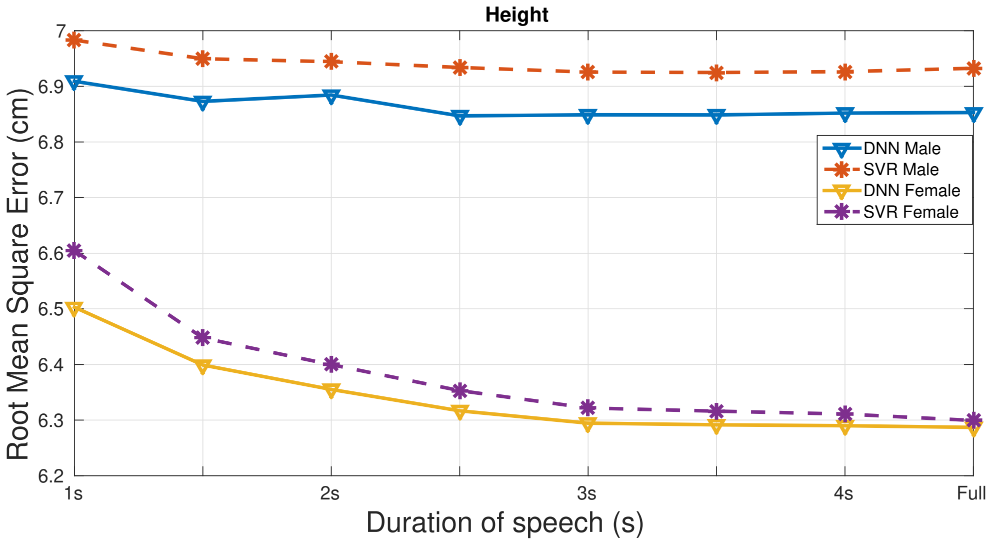

Figure 5.2: RMSE of height prediction using different lengths of speech data of both male and female speakers

error improves as expected and reaches a saturation around 3s. The GMM-UBM-SVR system (Babu and Vijayasenan (2017)) has a higher RMSE error consistently compared to the proposed joint model.

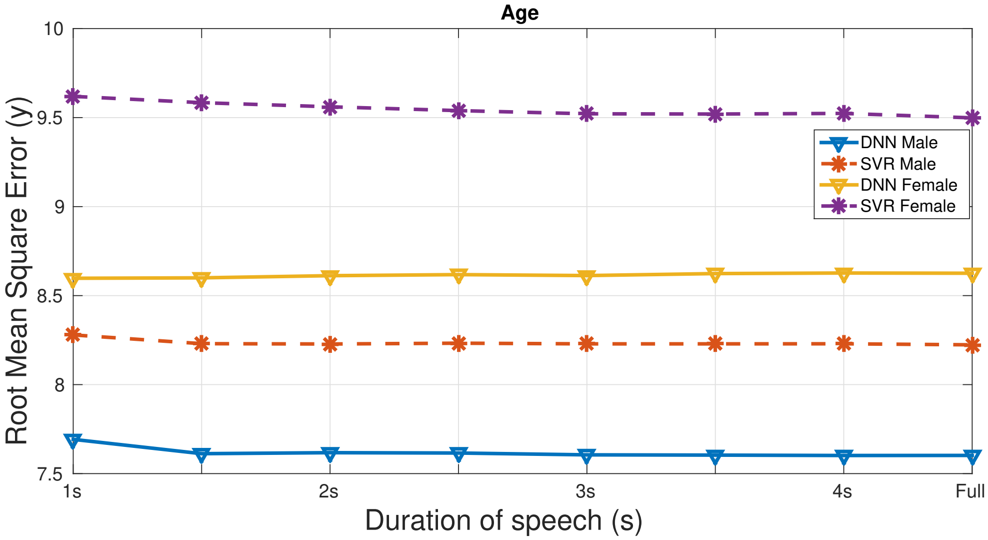

Figure 5.3: RMSE of age prediction using different lengths of speech data of both male and female speakers

The corresponding variations for age prediction is shown in Figure. 5.3. With only 1s speech available for prediction, the DNN model degrades only by 1.2% and 0.3% for male and female speakers. The RMSE of the GMM-UBM-SVR system is consistently more than the DNN system by 0.6 years for male speakers and 1 year for female speakers. Again the performance measure saturates around 3 seconds for the DNN system.

5.3.5 Error Analysis

In order to understand the errors, RMSE for height/age prediction is computed across different bins in the target values. Table 5.3 lists the results. In the training data, the height distribution is somewhat Gaussian shaped with lesser training data available for height values far away from the mean. In the results (Table 5.3), it can be noted that the height prediction RMSE is very high for the two extreme bins where the number of speakers are less as compared to the centre bins. However, in the case of age, the training data has a more uniform distribution, and it can be observed from Table 5.3 that the RMSE values do not change as much as in the case of height prediction. We hypothesize that height prediction can be further improved with a more uniform training data distribution.

Table 5.3: RMSE values of test speakers for different bins

	Height (cm)			
Range	MALE		FEMALE	
	# Train spkrs	Test	# Train spkrs	Test
$h < 150$	–	–	2	–
$150 < h < 160$	2	–	20	10.84
$160 < h < 170$	15	12.49	75	2.92
$170 < h < 180$	137	5.76	35	7.17
$180 < h < 190$	140	3.64	3	14.80
$190 < h$	32	12.98	–	–
	Age (years)			
Range	MALE		FEMALE	
	# Train spkrs	Test	# Train spkrs	Test
$a < 25$	67	7.54	47	6.70
$25 < a < 30$	132	6.21	46	5.11
$30 < a < 35$	66	6.88	14	5.95
$35 < a < 40$	28	6.65	9	7.45
$40 < a < 45$	13	9.67	9	3.80
$45 < a$	20	5.98	10	8.74

5.4 Experiments and Results Using AFDS & NISP Datasets

The physical parameter estimation is extended to shoulder size, waist size, weight along with the height and age of a speaker using the proposed DNN architecture. This DNN architecture jointly predicts all the physical parameters of the collected AFDS and NISP datasets. The same training and test splits of AFDS and NISP datasets (refer Section 4.5, Section 4.6 respectively) are used for the following experiments. All the experiments on AFDS and NSIP datasets are performed separately on the proposed DNN architecture.

As explained in Section 5.3.2, the first part of the DNN is initialized separately by training the model with the posteriors of GMM-UBM as target values. The first layer of the network is initialized with means of TIMIT UBM. The model is trained on training splits of both AFDS and NISP datasets. Due to less variability in the speakers, we adapted the TIMIT UBM means as well as speakers from AFDS and NISP datasets training splits to train the first part of the DNN.

The first order statistics of each utterance are computed at part-2 stage of the DNN model for both AFDS and NISP datasets. The frame level posteriors are computed by using sixty MFCC features (20 MFCC features along with 20 delta and 20 double deltas) for both datasets at L4 layer of the DNN. The utterance level Fstats are computed by taking the average of all the frames over the time at L5 layer and flatten at layer L6. This results in obtaining the first order statistics of dimension $60 \times 256 = 15360$.

For the sanity check the computed first order statistics from the DNN (without training the DNN) are fed to SVR for predicting each physical parameter separately for both the datasets. We denote this system as DNN-postr-SVR. The error metrics of each physical parameter using DNN-postr-SVR on AFDS and NISP datasets are tabulated in Table 5.4 and Table 5.5 respectively. It is observed that DNN posteriors are attaining the similar performance measure as our baseline GMM-UBM-SVR.

The DNN part-3, has a fully connected layer with five outputs (height, shoulder size, waist size, weight, and age) for AFDS and four outputs (height, shoulder size, weight, and age) for NISP dataset. This fully connected layer of the network is initialized with weights of the individual linear support vector regression algorithms. The male (female) neural network model is initialized from the male (female) SVR weights of each physical parameter (height, shoulder size, waist size, weight, and age) prediction.

The neural network is fully trained on 1-second segments of multilingual and multi accent speech data separately with male and female utterances of AFDS and NISP datasets' training data with respective targets. The physical parameters are jointly predicted for both datasets separately. The trained models are tested with the variable lengths of test data to predict the physical parameters (height, shoulder size, waist size, weight, and age) of the speaker. We denote this system as DNN-var-Pred. The error metrics of the predicted targets of AFDS are tabulated in Table 5.4. As a sanity check of the trained network, the RMSE values are compared with the RMSE values

Table 5.4: Comparison of RMSE values of physical parameters with estimation algorithms on AFDS dataset

Physical Parameter	TMP	GMM-UBM-SVR	DNN-postr-SVR	DNN-var-pred
		Male Speakers		
Height (cm)	6.91	6.54	6.57	**6.37**
Shoulder size (cm)	2.49	**2.41**	2.40	2.46
Waist size (cm)	7.34	7.16	7.15	**7.07**
Weight (kg)	9.59	9.33	9.33	**9.17**
Age (y)	3.41	**3.32**	3.34	3.33
Overall	14.54	14.07	14.08	**13.85**
		Female Speakers		
Height (cm)	5.94	**5.78**	5.78	5.97
Shoulder size (cm)	2.93	3.05	3.06	**2.92**
Waist size (cm)	7.19	7.12	7.08	**6.88**
Weight (kg)	8.42	8.37	8.36	**8.06**
Age (y)	1.03	1.05	**1.02**	1.21
Overall	12.94	12.83	12.8	**12.57**

of target mean predictor, as well as baseline RMSE values. Similarly, for NISP dataset the RMSE values are given in Table 5.5.

The proposed DNN architecture can predict the physical parameters jointly with an improvement over the TMP of each parameter in AFDS except age and height in female speakers. The DNN-var-Pred experiment shows the RMSE improvement of 7.8% over TMP in height, 3.7% in waist size, 4.4% in weight, 1.2% in shoulder size and 2.4% improvement in age of Male speakers in AFDS. Whereas for female speakers, there is an improvement of 4.3% in waist and weight and very minimal improvement in shoulder size estimation. There is no performance progress in height and age prediction. The system has degraded more in age estimation, as the majority of the female speakers are falling in mean age group and less number of speakers as well. The overall system performance of the DNN has improved by 4.86% for male speakers and 2.9% for female speakers when compared with TMP.

Similarly, in the case of male speakers of NISP dataset, there is an improvement of shoulder size, weight and age of the speakers by 1%, 6% and 2% respectively. And in the case of female speakers, the improvement in height is 2.6% and 10% in weight estimations over the TMP. TMP values are less than the predicted RMSE values for

Table 5.5: Comparison of RMSE values of physical parameters with estimation algorithms on NISP dataset

Physical Parameter	TMP	GMM-UBM-SVR	DNN-postr-SVR	DNN-var-pred
		Male Speakers		
Height (cm)	**6.17**	6.21	6.24	6.38
Shoulder size (cm)	2.58	2.67	2.57	**2.56**
Weight (kg)	9.57	9.38	8.92	**9.02**
Age (y)	5.60	5.69	5.65	**5.50**
Overall	12.95	12.89	12.53	**12.60**
		Female Speakers		
Height (cm)	6.93	7.02	6.79	**6.75**
Shoulder size (cm)	**3.52**	3.61	3.67	3.54
Weight (kg)	9.76	10.51	8.93	**8.77**
Age (y)	**5.57**	6.16	6.15	5.97
Overall	13.66	14.52	13.31	**13.06**

females shoulder size and age estimations. The overall performance improvement of the DNN is 2.7% and 4.5% for male and female speakers respectively when compared with the TMP.

5.5 Summary

In this work, we have proposed a deep neural network architecture to jointly predict speaker physical parameters from short-duration speech segments of all the three datasets (TIMIT, AFDS and NISP datasets). The neural network is initialized in a novel way using a conventional feature extraction (GMM-UBM super-vectors) and regression (SVR) scheme to avoid the requirement of a large amount of data. The network is trained with mean square error criterion and the joint model is able to improve the RMSE predictions.

The system is able to improve the RMSE of age prediction by more than 0.6 years, without degrading the RMSE for height prediction on TIMIT dataset. Analysis of shorter durations of speech reveals that the network only degrades around 3% at most with only 1 second of the speech input. Also, the performance saturates around 3seconds. The age prediction RMSE is lower than what is reported in literature

(Singh *et al.* (2016*b*)) that used stand-alone age prediction. The system performance is similar to full length speech data even with 3 seconds of speech for height and age estimation tasks.

The DNN system is able to jointly predict the all the physical parameters (height, age, shoulder size, waist size and weight) in a multilingual and multi accent setting. There is a consistent improvement in all physical parameter estimations in both the genders of AFDS except female speakers' height and age. The overall improvement of the system is by 2.9% and 4.8% for female and male speakers when compared with TMP. In the NISP dataset, the male the proposed system performed well for shoulder size, weight and age estimations in male speakers, and height and weight estimations in female speakers. The overall system performance has improved by 4.5% in female speakers and 2.7% in male speakers when compared with TMP.

In summary, the chapter's main contribution is the development of a joint model (DNN) for multiple physical parameters prediction, which is initialized in a novel way that enables the model to perform well on short duration speech utterances.

Chapter 6

Conclusions

The thesis's objective was to estimate the multiple physical parameters with short duration of speech data with out phone level transcriptions. This thesis also addresses for the physical parameter estimation in a multilingual and multi-accent setting.

The short-term Mel cepstral features, formants, and harmonics features are explored to estimate the physical parameters. These features are extracted at the utterance level that does not require phone level transcriptions. This common set of features are used in predicting the height and age of a speaker from the speech data. The combination of individual predicted targets using these features has improved the performance of physical parameter estimation system. The physical parameter estimation system is trained and tested on the standard monolingual TIMIT speech database (English) with a common feature set. It can predict the speaker's height and age with a minimum of 2 seconds of speech data, resulting in the state of the art results.

The proposed common set of features are extended to predict the multiple physical parameters (height, age, shoulder size, waist size, and weight) in a multilingual and multi-accent setting. Two multilingual and multi-accent speech datasets (AFDS and NISP datasets) have been created and the physical parameter details, linguistic, and geographical details are collected. The prediction error is also gets saturated at 2 seconds of speech when both the male and female speakers are considered in the multilingual and multi-accent setting. The prediction system is evaluated in matched and mismatched conditions to know how the system will degrade when the system is trained and tested on multiple languages. The analysis shown that, the system degrades utmost of 4% in males and 6.5% in female speakers of AFDS and utmost degradation of 10% in case of NISP dataset for the male and female speakers. This is the first attempt to predict the all (height, age, shoulder size, waist size, and weight) the physical parameters with the same common set of features.

A Deep neural network architecture is proposed for joint prediction of all the physical parameters in monolingual and multilingual settings using TIMIT, AFDS, and NISP datasets. The novel initialization scheme is proposed by using the conventional feature extraction (GMM-UBM super-vectors) and regression (SVR) scheme to avoid the requirement of a large amount of data. The network is trained with a mean square error criterion, and the joint model is able to improve the RMSE predictions. Analysis of shorter durations of speech reveals that the network only degrades around 3% at most with only 1 second of the speech input. Also, the performance saturates around 3seconds in predicting the height and age of a speaker using the TIMIT dataset. In the case of a multilingual setting using collected datasets, the predicted error metrics are less than the default predictor except in female age predictor in both AFDS and NISP datasets. In case of male speakers, the system performance is less than the default predictor in height estimation, and shoulder size and age estimations in female speakers of the NISP dataset.

Future Directions

The physical parameter system can be extended to language identification as well as accent detection which could help further more accurate identification of a speaker in case of forensic applications.

Integrating this physical parameter system along with language and accent in the end to end approach could be a potential advantage to the speaker profiling applications where the amount of available data is less.

Bibliography

Arsikere, H., **G. K. Leung**, **S. M. Lulich**, and **A. Alwan**, Automatic height estimation using the second subglottal resonance. *In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012.

Arsikere, H., **G. K. Leung**, **S. M. Lulich**, and **A. Alwan** (2013*a*). Automatic estimation of the first three subglottal resonances from adults speech signals with application to speaker height estimation. *Speech Communication*, **55**(1), 51–70.

Arsikere, H., **S. M. Lulich**, and **A. Alwan** (2011). Automatic estimation of the first subglottal resonance. *The Journal of the Acoustical Society of America*, **129**(5), EL197–EL203.

Arsikere, H., **S. M. Lulich**, and **A. Alwan** (2013*b*). Estimating speaker height and subglottal resonances using mfccs and gmms. *IEEE Signal Processing Letters*, **21**(2), 159–162.

Arsikere, H., **S. M. Lulich**, and **A. Alwan** (2014). Estimating speaker height and subglottal resonances using mfccs and gmms. *Signal Processing Letters, IEEE*, **21**(2), 159–162.

Babu, K. S. and **D. Vijayasenan**, Robust features for automatic estimation of physical parameters from speech. *In Region 10 Conference, TENCON 2017*. IEEE, 2017.

Bahari, M. H., **M. McLaren**, **H. Van hamme**, and **D. v. Leeuwen**, Age estimation from telephone speech using i-vectors. *In Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

Bocklet, T., **G. Stemmer**, **V. Zeissler**, and **E. Nöth**, Age and gender recognition based on multiple systems-early vs. late fusion. *In Eleventh Annual Conference of the International Speech Communication Association*. 2010.

Campbell, W. M., **D. E. Sturim**, and **D. A. Reynolds** (2006). Support vector machines using gmm supervectors for speaker verification. *IEEE signal processing letters*, **13**(5), 308–311.

Chollet, F. *et al.* (2015). Keras. https://keras.io.

Cieri, C., **D. Miller**, and **K. Walker**, The fisher corpus: a resource for the next generations of speech-to-text. *In LREC*, volume 4. 2004.

Collins, S. A. (2000). Men's voices and women's choices. *Animal behaviour*, **60**(6), 773–780.

Dusan, S., Estimation of speaker's height and vocal tract length from speech signal. *In Ninth European Conference on Speech Communication and Technology*. 2005.

DArcy, S. M., **M. J. Russell**, **S. R. Browning**, and **M. J. Tomlinson** (2004). The accents of the british isles (abi) corpus. *Proceedings Modélisations pour lIdentification des Langues*, 115–119.

Evans, S., **N. Neave**, and **D. Wakelin** (2006). Relationships between vocal characteristics and body size and shape in human males: an evolutionary explanation for a deep male voice. *Biological psychology*, **72**(2), 160–163.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America*, **102**(2), 1213–1222.

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

Ganchev, T., **I. Mporas**, and **N. Fakotakis**, Audio features selection for automatic height estimation from speech. *In Hellenic Conference on Artificial Intelligence*. Springer, 2010*a*.

Ganchev, T., **I. Mporas**, and **N. Fakotakis**, Automatic height estimation from speech in real-world setup. *In 2010 18th European Signal Processing Conference*. IEEE, 2010*b*.

Garofolo, J. S., **L. F. Lamel**, **W. M. Fisher**, **J. G. Fiscus**, and **D. S. Pallett** (1993). Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, **93**.

GermanSpeechDat(II) (). URL https://catalogue.elra.info/en-us/repository/browse/ELRA-S0096.

Ghahremani, P., **P. S. Nidadavolu**, **N. Chen**, **J. Villalba**, **D. Povey**, **S. Khudanpur**, and **N. Dehak** (2018). End-to-end deep neural network age estimation. *Proc. Interspeech 2018*, 277–281.

Gonzalez, J. (2003). Estimation of speakers' weight and height from speech: A reanalysis of data from multiple studies by lass and colleagues. *Perceptual and motor skills*, **96**(1), 297–304.

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

Gonzalez, S. and **M. Brookes** (2014). Pefac-a pitch estimation algorithm robust to high levels of noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**(2), 518–530.

Greisbach, R. (2007). Estimation of speaker height from formant frequencies. *International Journal of Speech Language and the Law*, **6**(2), 265–277.

Hansen, J. H., **K. Williams**, and **H. Bořil** (2015). Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models. *The Journal of the Acoustical Society of America*, **138**(2), 1052–1067.

Harper, M. (2013). The BABEL program and low resource speech technology. *Proc. of ASRU 2013*.

Jain, A. K., **A. Ross**, **S. Prabhakar**, *et al.* (2004). An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, **14**(1).

Kalluri, S. B., **A. Vijayakumar**, **D. Vijayasenan**, and **R. Singh**, Estimating multiple physical parameters from speech data. *In Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*. IEEE, 2016.

Krauss, R. M., **R. Freyberg**, and **E. Morsella** (2002). Inferring speakers physical attributes from their voices. *Journal of Experimental Social Psychology*, **38**(6), 618–625.

Lander, T. (). CSLU: Foreign Accented English release 1.2". URL https://catalog.ldc.upenn.edu/LDC2007S08.

Lass, N. J. and **W. S. Brown** (1978). Correlational study of speakers heights, weights, body surface areas, and speaking fundamental frequencies. *The Journal of the Acoustical Society of America*, **63**(4), 1218–1220.

Lass, N. J., **K. A. Scherbick**, **S. L. Davies**, and **T. D. Czarnecki** (1982). Effect of vocal disguise on estimations of speakers' heights and weights. *Perceptual and motor skills*, **54**(2), 643–649.

Layer, J. and **P. Truddgill** (1979). Phonetic and linguistic markers in speech. *Social Markers in Speech, Cambridge, CUP*, (1-C), 1–32.

Lehman, J. F. and **R. Singh**, Estimation of children's physical characteristics from their voices. *In INTERSPEECH*. 2016.

Li, M., **K. J. Han**, and **S. Narayanan** (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, **27**(1), 151–167.

Li, M., **C.-S. Jung**, and **K. J. Han**, Combining five acoustic level modeling methods for automatic speaker age and gender recognition. *In Eleventh Annual Conference of the International Speech Communication Association*. 2010.

Martin, A. F. and **C. S. Greenberg**, Nist 2008 speaker recognition evaluation: Performance across telephone and room microphone channels. *In Tenth Annual Conference of the International Speech Communication Association*. 2009.

Martin, A. F. and **C. S. Greenberg**, The nist 2010 speaker recognition evaluation. *In Eleventh Annual Conference of the International Speech Communication Association*. 2010.

Maxine Eskenazi, D. G., Jack Mostow (). The CMU Kids Corpus ". URL https://catalog.ldc.upenn.edu/LDC97S63.

Metze, F., **J. Ajmera**, **R. Englert**, **U. Bub**, **F. Burkhardt**, **J. Stegmann**, **C. Muller**, **R. Huber**, **B. Andrassy**, **J. G. Bauer**, *et al.*, Comparison of four approaches to age and gender recognition for telephone applications. *In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4. IEEE, 2007.

Mporas, I. and **T. Ganchev** (2009). Estimation of unknown speakers height from speech. *International Journal of Speech Technology*, **12**(4), 149–160.

Müller, C., Automatic recognition of speakers' age and gender on the basis of empirical studies. *In Ninth International Conference on Spoken Language Processing*. 2006.

Müller, C. and **F. Burkhardt**, Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age. *In Eighth Annual Conference of the International Speech Communication Association*. 2007.

Necioglu, B. F., **M. A. Clements**, and **T. P. Barnwell**, Unsupervised estimation of the human vocal tract length over sentence level utterances. *In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3. IEEE, 2000.

NIST-SRE (). NIST speaker recognition evaluation (SRE) series". URL https://www.nist.gov/itl/iad/mig/speaker-recognition.

Nolan, F., Forensic speaker identification and the phonetic description of voice quality. *In A figure of speech: A festschrift for John Laver*. Psychology Press, 2005, 385–411.

Pellom, B. L. and **J. H. Hansen**, Voice analysis in adverse conditions: the centennial olympic park bombing 911 call. *In Proceedings of 40th Midwest Symposium on Circuits and Systems. Dedicated to the Memory of Professor Mac Van Valkenburg*, volume 2. IEEE, 1997.

Peskin, B., **J. Navratil**, **J. Abramson**, **D. Jones**, **D. Klusacek**, **D. A. Reynolds**, and **B. Xiang**, Using prosodic and conversational features for high-performance speaker recognition: Report from jhu ws'02. *In Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 4. IEEE, 2003.

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

Poorjam, A. H., **M. H. Bahari**, **V. 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)*. IEEE, 2015.

Poorjam, A. H., **M. H. Bahari**, *et al.*, Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. *In Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*. IEEE, 2014.

Puts, D. A., **C. L. Apicella**, and **R. A. Cárdenas** (2012). Masculine voices signal men's threat potential in forager and industrial societies. *Proceedings of the Royal Society B: Biological Sciences*, **279**(1728), 601–609.

Reby, D. and **K. McComb** (2003). Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. *Animal behaviour*, **65**(3), 519–530.

Rendall, D., **S. Kollias**, **C. Ney**, and **P. Lloyd** (2005). Pitch (f 0) and formant profiles of human vowels and vowel-like baboon grunts: The role of vocalizer body size and voice-acoustic allometry. *The Journal of the Acoustical Society of America*, **117**(2), 944–955.

Reynolds, D. A., An overview of automatic speaker recognition technology. *In 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4. IEEE, 2002.

Sadjadi, S. O., **S. Ganapathy**, and **J. W. Pelecanos**, Speaker age estimation on conversational telephone speech using senone posterior based i-vectors. *In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.

Schötz, S., Acoustic analysis of adult speaker age. *In Speaker Classification I*. Springer, 2007, 88–107.

Schötz, S. and **C. Müller**, A study of acoustic correlates of speaker age. *In Speaker Classification II*. Springer, 2007, 1–9.

Schuller, B., **S. Steidl**, **A. Batliner**, **F. Burkhardt**, **L. Devillers**, **C. MüLler**, and **S. Narayanan** (2013). Paralinguistics in speech and languagestate-of-the-art and the challenge. *Computer Speech & Language*, **27**(1), 4–39.

Shivakumar, P. G., **M. Li**, **V. Dhandhania**, and **S. S. Narayanan**, Simplified and supervised i-vector modeling for speaker age regression. *In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

Singh, R., **J. Keshet**, and **E. Hovy**, Profiling hoax callers. *In 2016 IEEE Symposium on Technologies for Homeland Security (HST)*. IEEE, 2016*a*.

Singh, R., **B. Raj**, and **J. Baker**, Short-term analysis for estimating physical parameters of speakers. *In 2016 4th International Conference on Biometrics and Forensics (IWBF)*. IEEE, 2016*b*.

Smola, A. J. and **B. Schölkopf** (2004). A tutorial on support vector regression. *Statistics and computing*, **14**(3), 199–222.

Souza, L. B. R. D. and **M. M. D. Santos** (2018). Body mass index and acoustic voice parameters: is there a relationship? *Brazilian journal of otorhinolaryngology*, **84**(4), 410–415.

Spiegl, W., **G. Stemmer**, **E. Lasarcyk**, **V. Kolhatkar**, **A. Cassidy**, **B. Potard**, **S. Shum**, **Y. C. Song**, **P. Xu**, **P. Beyerlein**, *et al.*, Analyzing features for automatic age estimation on cross-sectional data. *In Tenth Annual Conference of the International Speech Communication Association*. 2009.

Tan, Z.-H. and **B. Lindberg** (2010). Low-complexity variable frame rate analysis for speech recognition and voice activity detection. *IEEE Journal of Selected Topics in Signal Processing*, **4**(5), 798–807.

Tanner, D. C. and **M. E. Tanner**, *Forensic aspects of speech patterns: voice prints, speaker profiling, lie and intoxication detection*. Lawyers & Judges Publishing Company, 2004.

Van Dommelen, W. A. and **B. H. Moxness** (1995). Acoustic parameters in speaker height and weight identification: sex-specific behaviour. *Language and speech*, **38**(3), 267–287.

van Heerden, C., **E. Barnard**, **M. Davel**, **C. van der Walt**, **E. van Dyk**, **M. Feld**, and **C. Müller**, Combining regression and classification methods for improving automatic speaker age recognition. *In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

Walker, K. and **S. Strassel**, The rats radio traffic collection system. *In Odyssey*. 2012.

Williams, K. A. and **J. H. Hansen**, Speaker height estimation combining gmm and linear regression subsystems. *In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.

Zazo, R., **P. S. Nidadavolu**, **N. Chen**, **J. Gonzalez-Rodriguez**, and **N. Dehak** (2018). Age estimation in short speech utterances based on lstm recurrent neural networks. *IEEE Access*, **6**, 22524–22530.

Publications Based on the Thesis

Journals :

1. **Shareef Babu Kalluri**, Deepu Vijayasenan, Sriram Ganapathy. **"Automatic Speaker Profiling from Short Duration Speech Data".** *Speech Communications*, vol.121, pg.16-28, May 2020. (Elsevier)
   DOI : 10.1016/j.specom.2020.03.008

Conferences :

1. **Kalluri, Shareef Babu** , Deepu Vijayasenan, Sriram Ganapathy, Ragesh Rajan M, Prashant Krishnan, **"NISP: A Multi-lingual Multi-accent Dataset for Speaker Profiling"**,in the proceedings of the *46th IEEE International Conference on Acoustics, Speech and Signal Processing (**ICASSP**)*. Toronto, Ontario, Canada, IEEE, 2021.
   DOI: 10.1109/ICASSP39728.2021.9414349

2. **Kalluri, Shareef Babu** , Deepu Vijayasenan, Sriram Ganapathy "A Deep Neural Network based End to End Model for Joint Height and Age Estimation from Short Duration Speech" in the proceedings of *44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
   DOI: 10.1109/ICASSP.2019.8683397

3. **Babu, Kalluri Shareef**, and Deepu Vijayasenan. "Robust Features for Automatic Estimation of Physical Parameters from Speech", in the proceedings of *TENCON 2017-2017 IEEE Region 10 Conference* , IEEE, 2017.
   DOI: 10.1109/TENCON.2017.8228097

4. **Kalluri, Shareef Babu**, Ashwin Vijayakumar, Deepu Vijayasenan, and Rita Singh. "Estimating multiple physical parameters from speech data." in the proceedings of *IEEE 26th International Workshop on Machine Learning for Signal*

Processing (MLSP). IEEE, 2016.
DOI: 10.1109/MLSP.2016.7738873

5. Vijayasenan, Deepu, **Shareef Babu Kalluri**, K. Sreekanth, and Ansal Issac. "Study of Wireless Channel Effects on Audio Forensics" in the proceedings of *22nd Annual International Conference on Advanced Computing and Communication (ADCOM)*, pp. 33-37. IEEE, 2016.
DOI: 10.1109/ADCOM.2016.15

Other publications :

1. Kotra Venkata Sai Ritwik, **Kalluri, Shareef Babu**, Deepu Vijayasenan, "COVID-19 Detection from Spectral features on the DiCOVA Dataset", in the proceedings of *Interspeech 2021*, Brno, Czech Republic.

2. Kotra Venkata Sai Ritwik, **Kalluri, Shareef Babu**, Deepu Vijayasenan, "COVID-19 Patient Detection from Telephone Quality Speech Data", arXiv preprint arXiv:2011.04299 (2020).

Bio-data

NAME : Kalluri Shareef Babu

CONTACT DETAILS

Address	: S/o K Babji, #7/29,Cross Roads , Kalikiri, Kalikiri (P&M), Chittoor Dist., Andhra Pradesh–517234
☎	: 9620789927
🖂	: shareefbabu1@gmail.com

EDUCATIONAL QUALIFICATIONS

Doctor of Philosophy (Ph.D)

National Institute of Technology Karnataka, Surathkal Dec 2013–Dec 2020

Master of Technology (M.Tech)

Jawaharlal Nehru Technological University (JNTU)- Anantapur,
Andhra Pradesh. 2011-2013

Branch : Digital Electronics and Communication Systems

Bachelor of Technology (B.Tech)

Jawaharlal Nehru Technological University (JNTU)- Anantapur,
Andhra Pradesh., 2007-2011

Branch : Electronics and Communication Engineering

Research Interests

Speech Signal Processing, Speaker profiling, Pattern Recognition, Machine Learning and Deep learning.

I Joined as Post Doctoral Fellow at IUI-Lab, SEIKEI University, Japan.