Speech Communication 121 (2020) 16–28

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Automatic speaker profiling from short duration speech data

Shareef Babu Kalluri [a,*], Deepu Vijayasenan [a], Sriram Ganapathy [b]

[a] Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, Mangalore, Karnataka 575025, India
[b] Learning and Extraction of Acoustic Patterns (LEAP) Lab, Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

ARTICLE INFO

Keywords:
Speaker profiling
Short duration
Formants
Harmonics

ABSTRACT

Many paralinguistic applications of speech demand the extraction of information about the speaker characteristics from as little speech data as possible. In this work, we explore the estimation of multiple physical parameters of the speaker from the short duration of speech in a multilingual setting. We explore different feature streams for age and body build estimation derived from the speech spectrum at different resolutions, namely – short-term log-mel spectrogram, formant features and harmonic features of the speech. The statistics of these features over the speech recording 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 is able to achieve results that outperform previously published results in height and age estimation. Furthermore, the estimation errors from these different feature streams are complementary, which allows the combination of estimates from these feature streams to further improve the results. 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 of 5.2 years, and 5.6 years for male, and female speakers respectively. We also extend the same physical parameter estimation to other body build parameters like shoulder width, waist size and weight along with height on a dataset we collected for speaker profiling. The duration analysis of the proposed scheme shows that the state of the art results can be achieved using only around 1–2 s 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.

1. Introduction

Apart from the textual message, human speech contains information about speaker identity, emotion, gender, accent etc. The extraction of speaker traits (parameters) from the speech data could further aid in speaker identification systems as well as in speaker clustering and diarization systems. The main challenge in estimating any such information is the separation of linguistic content and speaker traits.

In this paper, we try to address the problem of estimating physical parameters from the short duration of speech in a multilingual setting. This involves in predicting speaker meta information such as age and parameters of body build like height, weight, shoulder size and waist size. 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., 2016a; 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). In case of age estimation, researches have focused to identify the age group of a speaker (children, youth, adult and senior) from speech for most of the commercial applications (targeted advertisements, caller-agent pairing in call-centers etc.) besides other applications like surveillance, forensics to narrow down on suspects from hoax/threat calls etc (Nolan, 2005; Singh et al., 2016a; Schuller et al., 2013).

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. For example, DARPA RATS program targeted development of speaker and language recognition technology with as little as 3 seconds (s) of speech (Walker and Strassel, 2012). Thus, development of speaker profiling methods in short duration audio is important.

1.1. 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

* Corresponding author.
E-mail address: shareefbabu1@gmail.com (S.B. Kalluri).

https://doi.org/10.1016/j.specom.2020.03.008
Received 19 March 2019; Received in revised form 11 March 2020; Accepted 31 March 2020
Available online 1 May 2020

(Layer and Trudgill, 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) (Fitch and Giedd, 1999). It is also worthwhile noting that the sample size in this study for adult subjects (17–25 years) was quite small (six female and 13 male).

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.

Previous attempts (Layer and Trudgill, 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). 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).

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 5 s. length is challenging yielding significantly worse results making the systems inoperable for realistic applications (Ghahremani et al., 2018). In this work, we address the problem of reliably extracting height/age information from short duration speech 2–3 s. segments without using any phonetic information. We also extend the work to estimating more physical parameters (shoulder size, waist size, and weight). The main novelty of the proposed work lies in developing a unified framework for height/age and other physical parameter estimation. This is achieved using features that extract spectral structure of speech signal in terms of format frequencies (peak locations in wide-band spectrum estimated using an autoregressive model) and harmonic frequency locations.

1.2. Organization of the paper

The rest of the paper is organized as follows. Section 2 describes about the speaker profiling literature, motivation to carry out this work and contributions of the paper. Section 3 briefs about datasets, features extracted and regression technique used to estimate the physical parameters using speech data. Section 4 describes about the experiments conducted to estimate height and age of a speaker in monolingual setting using TIMIT dataset. Also this section discuss about the experiments performed on multiple physical parameters on multilingual setting using AFDS dataset in Section 4.3. A duration analysis is performed to know the minimum amount of speech data required for estimating physical parameters and this is explained in Section 4.4. Finally, conclusions are presented in Section 5.

2. Speaker profiling literature

While there is information about 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.

2.1. 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 a input to system.

The researchers predict 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; 2013a) and fundamental frequency (Dusan, 2005). Phone specific (vowels like /iy/,/ae/,/ey/,/ih/,/eh/ etc.) short term features like (MFCC, LPC) and formants shows a correlation of around 0.75 and for F_0 it is 0.59 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., 2013a). 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.

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 6.36 cm for male and 6.8 cm 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 5.37 cm and 5.49 cm for male and female speakers respectively. Later line spectral frequencies were added to the feature set resulting in a lower MAE 4.93 cm and 4.76 cm 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). 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 5.3 cm and RMSE of 6.8 cm 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., 2013b). For example the MFCC features are modeled 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 5.7 cm for male and female speakers respectively (Arsikere et al., 2013b).

Singh et al. (2016b) 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), and Ganchev et al. (2010). 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., 2016b).

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 uses 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 – Multilayer Perceptron), 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 features are preceded by background GMM, SVM etc. for the age estimation (Li et al., 2013; van Heerden et al., 2010). With Interspeech 2010 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 and RMSE of 7.8 years for male, and for female speakers, MAE is 6.5 years and RMSE is 8.9 years on TIMIT dataset (Singh et al., 2016b).

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 correlation of male and female speakers are 8.72 years, 0.37, and 7.95 years, 0.54 respectively when 3 s 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 5 s chunks of speech data are 5.78 years, 0.74 for male, 4.23 years, 0.87 for female respectively (Ghahremani et al., 2018). Table 1 shows the summary of the prior works methods and features for height and age estimation tasks.

2.3. Other physical characteristics

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).

2.4. Limitations of prior work

Majority of the speaker profiling works of the past concentrate on estimating only one physical parameter – either age 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; 2013a; 2011; 2013b) 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 (> 10 s) (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) 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., 2016b). So far the only work that addressed both height and age estimation from short duration speech is Singh et al. (2016b). However, the prior work on short duration speech shows that dealing with utterances of < 5 s. of speech in physical parameter estimation is challenging. To the best of the authors' knowledge, literature does not address the physical parameter estimation from short duration multilingual speech data.

2.5. Contributions from this work

In this work, we attempt to address the main two challenges for physical trait estimation, one is short duration of utterances and second is the multilingual nature of the data. The goal is to come up with a common feature input for all physical parameter prediction systems. The proposed features do not require phone level transcriptions. We consider

Table 1
Summary of prior work in age and height estimation.

Literature summary on age			
Reference	Motivation	Features	Model
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
Mporas and Ganchev (2009), Ganchev et al. (2010)	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, 2013a, 2011, 2013b)	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. (2016b)	Forensics, target advertisements	Short term spectral features	Random Forest

different characteristics of the speech signal – short-term spectral features, fundamental frequency, formant frequency locations and narrowband speech harmonics. With many experimental results, we show that the proposed approach of using spectral features is useful in the prediction of height/age and other physical attributes of the speaker.

To the best of our knowledge, this paper presents the first work of its kind to illustrate the estimation of physical parameters from short durations of speech signal in a multilingual setup. We perform height and age estimation experiments in the TIMIT database (Garofolo et al., 1993) where the speech recordings are 2–3 s duration. The combination of these features attain the MAE of 5.2 years (male) and 5.6 years (female) in age estimation and in case of height estimation the MAE is of 5.2 cm for males and 4.8 cm for female speakers. In these experiments, the combination of proposed features shows significant improvements over the previously published results on the same dataset (Arsikere et al., 2013a; Singh et al., 2016b). We extend the same approach to multilingual setting to predict multiple physical parameters like shoulder width, waist size, weight along with height on a dataset. Finally, we investigate the minimum amount of speech required to perform physical parameter estimation on both TIMIT and AFDS datasets.

3. Methodology

In this work, we use two datasets for our experiments and analysis. One is the standard TIMIT dataset (Garofolo et al., 1993), and the second one is a multilingual dataset, Audio Forensic Dataset (AFDS) (Kalluri et al., 2016), collected for this purpose. We extract three different features which does not require the phoneme level transcriptions for short speech segments. The utterance level statistics of the extracted features is given to a support vector regression to estimate the physical parameters.

3.1. Datasets

The TIMIT dataset has 630 speakers, each speaker has contributed 10 recordings. Each of the ten recordings per speaker is considered as a separate input data sample. For training set, we have 462 speakers (326 male and 136 female speakers) and for testing 168 (56 female and 112 male speakers). The statistics of the dataset is given in Table 2. The training and validation splits has 4610 utterances which includes 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. Each input utterance had 1–3 s of speech data for height and age prediction.

Table 2
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 (*year*)	20.63	75.77	30.52	7.57
Female speakers				
Height (*cm*)	144.78	182.88	165.80	6.71
Age (*year*)	21.08	67.35	30.03	8.70
Male and female speakers				
Height (*cm*)	144.78	203.20	175.50	9.47
Age (*year*)	20.63	75.77	30.37	7.98

The second one is a dataset, collected from diverse dialects of individuals across India for this study. This dataset was named as Audio Forensic Dataset (AFDS) (Kalluri et al., 2016). This dataset contains the speaker details like height, weight, shoulder width, waist size along with the speech utterances. Speech is recorded at a sampling frequency of 16 kHz. Each speaker provided around 2 min of speech data in three sessions, with each session lasting around 40 s. This dataset is linguistically diverse with people having 12 different native languages. Each speaker is asked to read news articles in their native language as well as in English. This speech corpus contains 207 speakers including 161 males and 46 females. The speakers are in the age group of 18–37 years. The height, shoulder width and waist size are measured in centimeters (cm) and weight in kilograms (kg). The statistics of the dataset are tabulated in Table 3.

Table 3
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
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
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

For evaluation purpose, the dataset is divided into training and testing datasets. Training data has 137 speakers (951 utterances) consisting of 104 males (727 utterances) and 33 females (224 utterances). Testing data has 70 speakers (538 utterances) consisting of 57 males (434 utterances) and 13 females (104 utterances). Train and test splits includes both English and native language. Both the training and testing splits contain speakers across the 12 different languages. There is no overlap of speakers in both the datasets of training and test splits.

3.2. Feature extraction

In this paper, we try to come up with a common set of features that can be used for the physical parameters estimation. We explore different features which uncover the underlying the 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; Dusan, 2005; Poorjam et al., 2014). The fundamental and formant frequencies contain information about physical parameters of a speaker (Rendall et al., 2005; Hansen et al., 2015; Arsikere et al., 2013a). The narrowband spectral harmonics capture the fine spectral structure on a coarse temporal scale. The log harmonics are used in estimating the age and gender of a speaker (Li et al., 2013). We use both frequency and amplitude of the spectral peaks as harmonic features (to capture jitter and shimmer characteristics of speech).

3.2.1. Feature extraction using mel filter bank features and UBM

Mel filter cepstral coefficients (MFCC): The 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; Poorjam et al., 2014). In our work, we extract 20 mel frequency cepstral coefficients (using a window length of 25 ms with a shift of 10 ms) along with delta and double delta features (yielding 60 MFCC features).

Mel filter bank features: In our work, we use the logarithm of the mel spectral energy in short-term windows (25 ms with a shift of 10 ms) 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. We adopt a supervector (Reynolds, 2002) approach which can summarize the gross spectral changes in order to normalize the effect of phonetic information in the short-term spectral representation.

Statistical representation: In order to form a background UBM model, a Gaussian Mixture Model 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 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ represents the input MFCC feature vectors and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$ 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 :

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

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 jth 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:

$$\mathbf{f}_i^j = \mathbf{y}_i p(j|\mathbf{x}_i), \tag{2}$$

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

$$p(j|\mathbf{x}_i) = \frac{w_j \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}_j, \mathbf{C}_j)}{\sum_{j=1}^{M} w_j \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}_j, \mathbf{C}_j)}. \tag{3}$$

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:

$$\mathbf{F} = \frac{1}{T} \sum_{i=1}^{T} \mathbf{F}_i \tag{4}$$

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. These features are used in support vector regression to estimate the physical parameter.

3.2.2. Extraction of fundamental and formant frequency features

We compute the fundamental frequency from a wideband analysis of speech signal (temporal window size of 20 ms with a shift of 10 ms). 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 20 ms with a shift of 10 ms) 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 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 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 modest. 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),

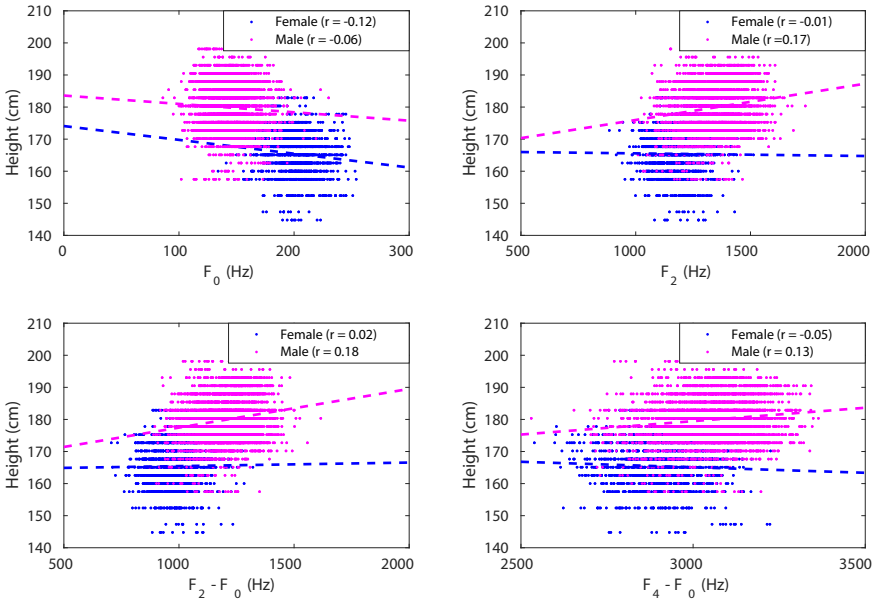


Fig. 1. Scatter plot of fundamental and formant frequency estimates with the speaker height for TIMIT training set. Value in the brackets shows the correlation (r) between formants and corresponding physical parameter (height) for male and female speakers. The best fit line is also shown for both male and female speakers separately.

difference between F_0 and formants (Rendall et al., 2005). For example, we find the correlations between difference of F_0 and formants ($F_1 - F_0, F_2 - F_0, F_3 - F_0, F_4 - F_0$), Fig. 1 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).

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.

3.2.3. Extraction of harmonic features

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. It has been shown that variations in frequency (jitter) and amplitude (shimmer) contain useful information about age as well (Müller and Burkhardt, 2007).

Using an AR model (order 80 with a window length of 60ms and a shift 10ms) of the spectrum, the peak locations (locations of the poles of the AR model) are identified. The logarithm of the frequency and amplitude of spectral peaks are computed at each frame. 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. Fig. 2 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 Fig. 3 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 frequency location features capture jitter features and amplitude features captures shimmer features.

3.3. Prediction using support vector regression

We use a standard support vector regression (SVR) (Smola and Vapnik, 1997) as the model for predicting the target of each physical parameter values given the input features. 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:

$$\begin{aligned} &\min \tfrac{1}{2}\mathbf{w^T w} \text{ subject to} \\ &|\mathbf{w^T y_i} + b - t_i| < \epsilon \end{aligned} \tag{5}$$

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 SVR optimization function aims to reduce the deviation from the target values by the parameter ϵ. We have also explored both linear and nonlinear kernels in this paper. In case of multiple features, we average the individual SVR outputs.

4. Experiments and results

We perform height and age estimation experiments on TIMIT dataset. We use the standard train and test split in TIMIT. The algorithms are benchmarked using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics.

$$\begin{aligned} MAE &= \tfrac{1}{N}\textstyle\sum_i |x_i^{pred} - x_i^{true}| \\ RMSE &= \sqrt{\tfrac{1}{N}\textstyle\sum_i (x_i^{pred} - x_i^{true})^2} \end{aligned} \tag{6}$$

where x_i^{pred} and x_i^{true} are the predicted and target values for the ith test utterance.

4.1. Results with individual features

In order to understand the effect of each feature separately, we evaluated the individual performance of the features. All hyper parameters

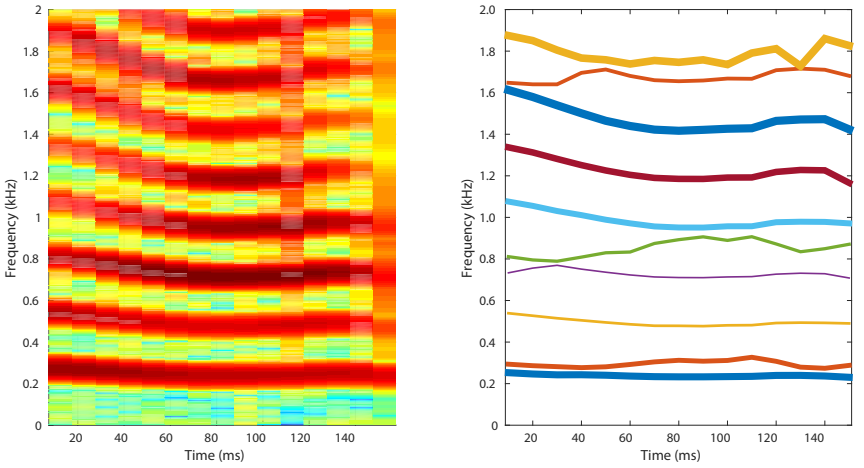


Fig. 2. Spectrogram for vowel /AE/ and corresponding trajectories of first 10 peaks locations in a narrow-band spectrogram estimated using an AR model.

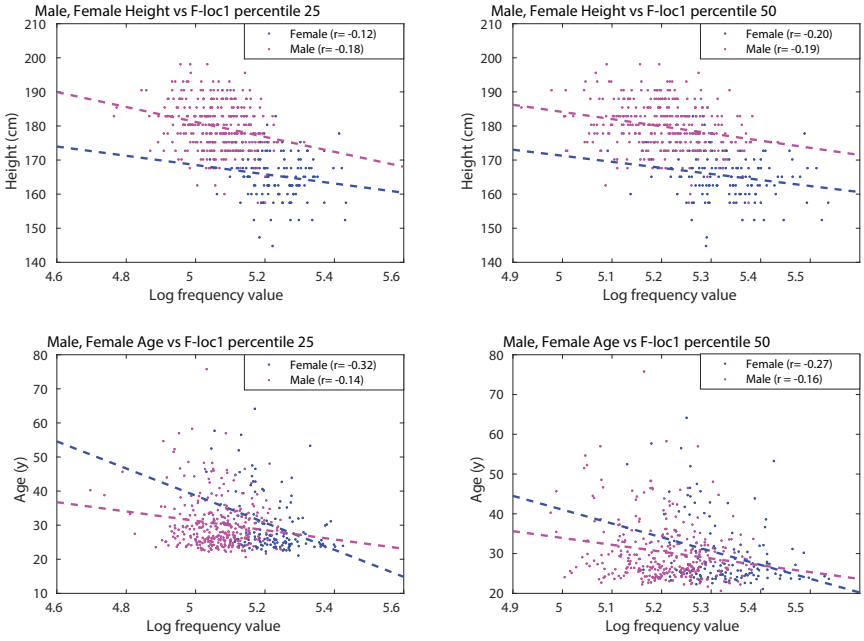


Fig. 3. Scatter plot of Harmonic percentiles (25 and 50) vs physical parameter (height and age) for male and female speakers of TIMIT training data. Correlation (r) value between harmonic percentile and physical parameters (height and age) is given in brackets for male and female speakers. The best fit line is also shown for both male and female speakers separately.

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. (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 18th order (fixed based on validation set) model is used with a 20 ms length window with 10 ms shift. The 5th, 25th, 50th, 75th and 95th 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 60 ms and shift 10 ms). 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. Fig. 4 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, the

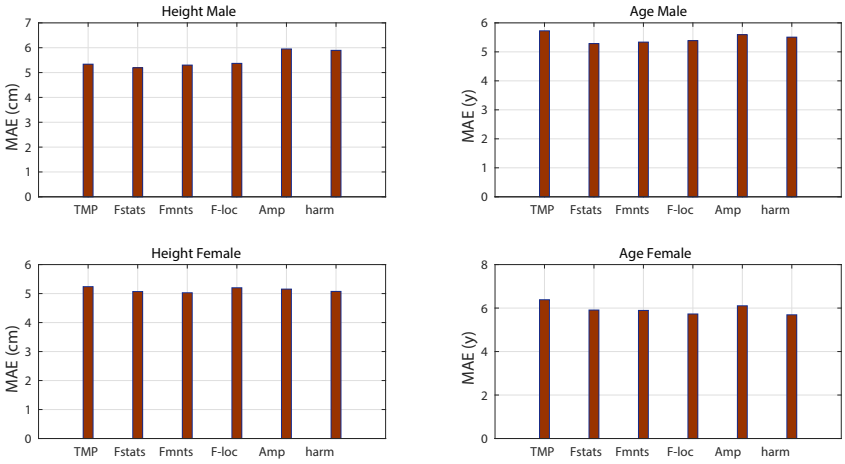


Fig. 4. Mean absolute error comparison with training mean predictor (TMP) and prediction of different systems using first order statistics (Fstats), formants (Fmnts), harmonic frequency locations (F-loc), amplitude (Amp) and harmonic features (harmonic frequency locations and amplitude features together: harm) for height (left side) and age (right side) estimation using the TIMIT dataset.

Table 4
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)	–	–	–	–	5.3	6.8
Arsikere et al. (2013a)	5.6	6.9	5.0	6.4	5.4	6.8
Singh et al. (2016b)	**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 (year) 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. (2016b)	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

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.

4.2. Results with feature combination

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 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). All our analyses use the standard TIMIT train and test splits. Table 4 reports the results along with the recent baseline which uses the standard train and test splits of TIMIT dataset (Singh et al., 2016b).

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,

Table 5
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	–	–

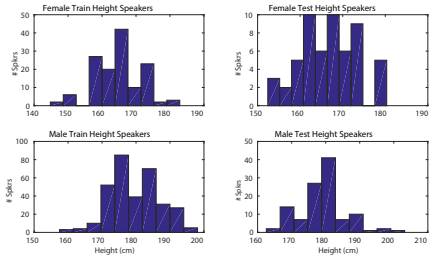
the relative improvement in height prediction of Comb-3 w.r.t to TMP is 2.94% and 6.15% for male and female speakers respectively. Similarly, for age prediction there is an 5.75% relative improvement for female speakers and no improvement for the male speakers.

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.

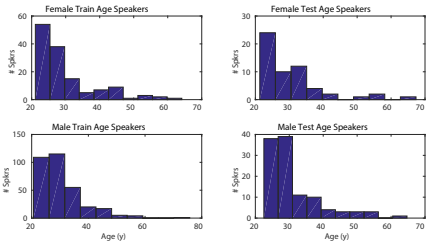
In case of height estimation, we also compare with three other baselines. The error metrics MAE and RMSE of the proposed systems as well as the baseline results are presented in Table 4. In case of female speakers both MAE and RMSE performances of Comb-3 are better than the baseline for height estimation. 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 5 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 and 6 for male speakers and 2 and 5 for female speakers) in Table 5. 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 Fig. 5(a). 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 in TIMIT is by Singh et al. (2016b). Comparison of this baseline with our results and TMP is presented in Table 4. 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



(a) Speaker Height – Training data (Left) and Test data (Right)



(b) Speaker Age – Training data (Left) and Test data (Right)

Fig. 5. Histogram of TIMIT dataset gender specific training data and test data – height and age.

Table 6
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.	$20 \le a < 25$	67	4.6	4.8	47	2.7	3.0
2.	$25 \le a < 30$	132	1.8	2.1	46	2.0	2.4
3.	$30 \le a < 35$	66	2.9	3.4	14	4.7	5.2
4.	$35 \le a < 40$	28	7.8	8.1	9	8.8	8.9
5.	$40 \le a < 45$	13	13.0	13.1	9	13.0	13.1
6.	$45 \le a < 55$	16	22.2	22.4	7	24.9	25.0
7.	$55 \le a < 65$	3	35.5	35.5	3	21.9	21.9
8.	$65 \le a < 76$	1	–	–	0	35.0	35.1

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 shown in Table 6. 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 6). All these age groups have very few training speakers. Therefore, the RMSE error in these three groups are large (greater than 22 years) and is dominates the overall RMSE performance. The histogram of gender specific speaker age in both training and testing datasets are depicted in Fig. 5(b). It can be seen that there are very few number of speakers above 45 years in training.

4.3. Extension to other physical parameters

We extend the same approach followed to estimate height and age to more physical parameters in a multilingual setting using the AFDS dataset as described in Section 3.1. We have analysed the correlation of height with other parameters like shoulder size, waist size and weight on AFDS dataset. In the case of height the correlation values are small (0.2, 0.3 and 0.4 for shoulder size, waist size and weight respectively) for male speakers. The correlation values with age was negligible. Thus, these are parameters that cannot be predicted from height or age. We do not report results on only female data since, the number of female speakers is small.

In this regard, we use the same feature set (i.e., fundamental frequency, formants, harmonic features, and first order statistics of the speech signal) as explained in Section 3.2. In order to compute the first order statistics on AFDS, we have extracted 20 MFCCs along with deltas and double deltas and 40 filter bank features. We have used the GMM UBM learned from training data of TIMIT dataset itself, as the number of training utterances are less in AFDS. The Fstats are computed on AFDS using the Eq. (4) (refer to Section 3.2.1).

The fundamental frequency, formants and harmonic features are extracted from the AFDS speech data along with its percentiles as explained in Sections 3.2.2 and 3.2.3. These statistical features are fed to the support vector regression for the physical parameter estimation. The mean absolute error of each feature is compared with the training data mean predictor of each physical parameter (height, shoulder size, waist size and weight) is shown in Fig. 6. The Fstats and formants shows better MAE performance for all the physical parameters. The harmonic features are better than TMP in case of height estimation.

Simple averaging is then performed on the predicted test targets obtained from formant features, Fstats and harmonics features (refer to Section 4.2). The comparison of combination results and training data mean predictor are listed in Table 7. The table also lists an earlier algorithm developed by the authors as the baseline (Kalluri et al., 2016). All the results use the same train and test split described in Section 3.1 (same splits are used in our previous work Kalluri et al., 2016). The baseline performs support vector regression of a bag of words representation extracted from the short-term spectrum of the speech. The performance metrics both MAE and RMSE on Comb-3 are better than the baseline for all speakers (both male and female speakers) except in MAE of weight estimation. With Comb-3, there is a substantial improvement of MAE and RMSE in all the physical parameters when compared with the TMP when only male speakers are considered. For further analysis we use Comb-3 set of features.

4.4. Duration analysis

In order to analyze 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 0.25 s 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 Fig. 7.

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 0.5 s) 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 s. 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 1 s of speech data, when both male and females speakers are considered, the model is able to obtain prediction error MAE of 5.27 cm at par with Ganchev et al. (2010) 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

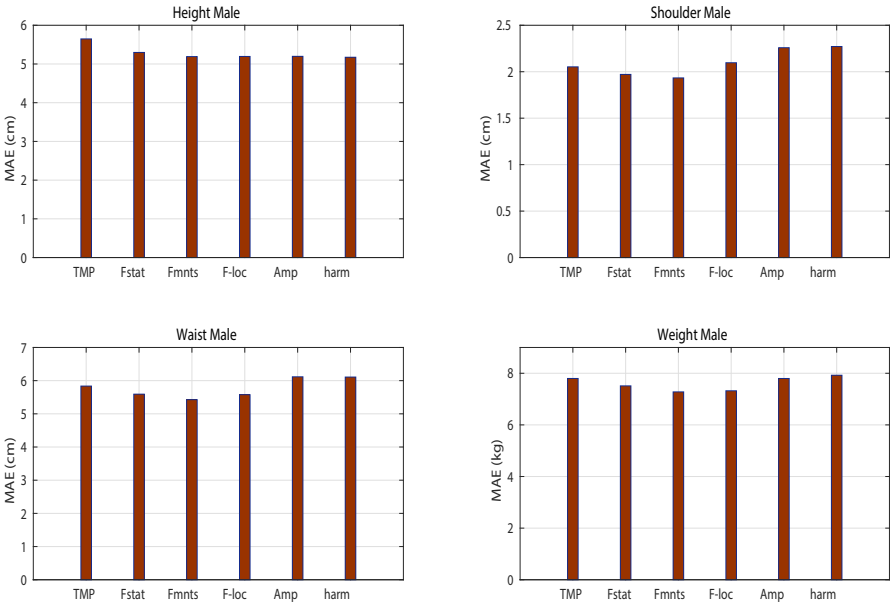


Fig. 6. Mean absolute error of male speakers compared with training mean predictor (TMP) and prediction of different features i.e, first order statistics (Fstats), formants (Fmnts), harmonic frequency locations (F-loc), amplitude (Amp) and harmonic features (harmonic frequency locations and amplitude features together: harm) of physical parameters (Height, Shoulder width, Waist and Weight) using AFDS.

Table 7
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 baseline results of AFDS.

Multiple physical parameter estimation – all (male + female)

	TMP		Baseline (Kalluri et al., 2016)		Comb-1		Comb-2		Comb-3	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Height (cm)	6.8	8.2	5.2	6.6	5.1	6.3	5.0	6.1	**5.0**	**6.1**
Shoulder (cm)	2.8	3.4	2.1	2.6	2.0	2.4	2.0	2.4	**1.9**	**2.4**
Waist (cm)	5.6	7.3	5.4	7.1	5.3	6.9	5.4	6.9	**5.5**	**7.0**
Weight (kg)	8.3	10.57	**6.7**	8.9	6.9	9.0	7.0	8.9	6.9	**8.8**

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**

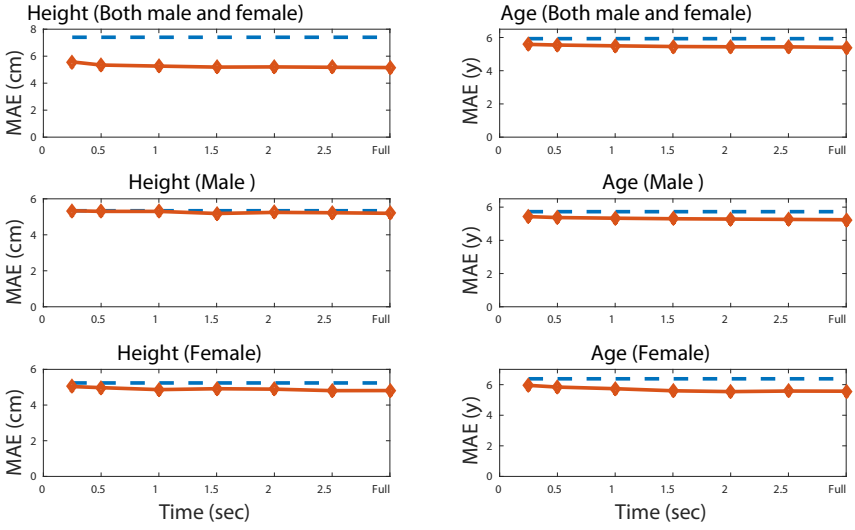


Fig. 7. MAE vs duration of utterance, for physical parameters' (height, age) estimation from TIMIT database. The horizontal dashed line represent training data mean predictor (TMP) benchmark.

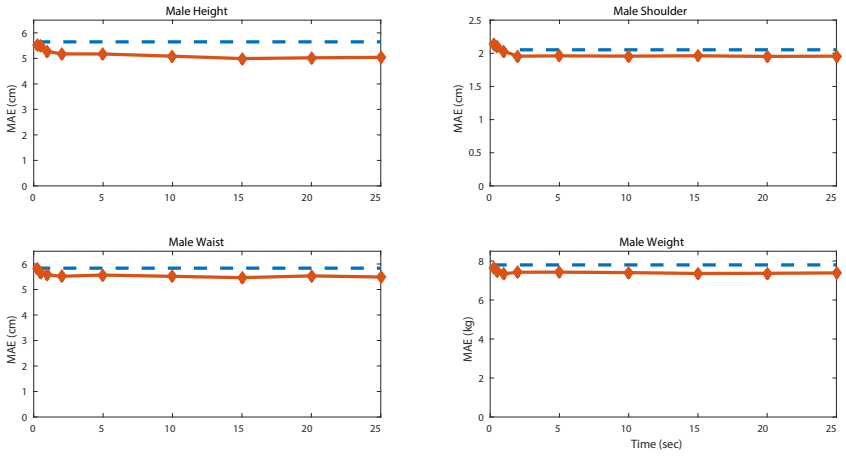


Fig. 8. MAE vs duration of utterance, for physical parameters' (height, shoulder width, waist size and weight) estimation from AFDS database. The horizontal dashed line represent training data mean predictor (TMP) benchmark.

speakers are considered together, the minimum duration of speech required to get the state-of-the-art prediction error is 0.5 s (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 Fig. 7. About, 2 s of speech data is required to get a performance comparable to the full length data.

We also extend the same duration analysis on other physical parameters (shoulder width, waist size and weight along with height) using the AFDS dataset. The system performance is evaluated for different lengths of speech files ranging from 0.25 s to full duration(around 40 s). We observed that mean absolute errors of each physical parameter for different durations' of speech signal is less than the training data mean prediction error except shoulder size by only using 0.5 s for male speakers. From this, it is evident that the system is reliably able to predict the physical parameters from 0.5 s duration of speech signal with prediction error less than the training data mean. The duration of the speech at which the prediction error saturates is around 2 s when both genders data is considered together. The mean absolute error for height is 5.1 cm, shoulder width is 1.9 cm, waist size is 5.4 cm and for weight is 6.9 kg when we have 2 s of speech data, where as when the available speech data is 40 s, we have 5.0 cm, 1.9 cm, 5.5 cm and 6.9 kg for height, shoulder width, waist size and weight respectively when both male and female speakers are considered together. The variation of MAE with respect to utterance duration for male speakers is shown in Fig. 8. For male speakers also the MAE saturates around 2 s as like above mentioned case (both male + female speakers). The change in MAE when full duration (40 s) and 2 s considered is 0.1 cm in height, and there is no change in MAE for other physical parameters like shoulder size, waist size and weight estimation.

4.5. Summary

In short, it can be seen that each of the physical parameter prediction error is less than the TMP even with short speech segments (around 0.5 s). We are able to achieve the state-of-the-art results with around 1–2 s for all the physical parameters. The MAE of the proposed height estimation system on TIMIT (5.2 cm for male, 4.8 cm for female) is similar to the best height estimation results (5.0 cm for male and 4.8 cm for female) (Hansen et al., 2015). Note that this system (Hansen et al., 2015) requires speech transcription for computing the phoneme specific features. In case of age estimation, the MAE of the proposed system (5.2 years for male and 5.6 years for female) is better than the state of the art result (MAE of 5.5 years for male, 6.5 years for female) reported on TIMIT (Singh et al., 2016b). Also, we demonstrate similar performances for other physical parameters in a multi-lingual setting. In summary, we hypothesize that the proposed methods could be used for speaker profiling where the duration of available speech data is limited.

5. Conclusions

In this work, we have explored the estimation of multiple physical parameters from short duration speech segments. In addition to conventional short-term spectral features, we also show that formant frequency features and harmonic structure of speech could be used as input to these tasks. Each of the individual features perform equally well on the test data and are able to achieve results that are comparable to state-of-the-art. 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 male, female and all speakers using the TIMIT dataset.

We have also presented the details of a new dataset where more speaker attributes like height, shoulder width, waist size and weight are collected. Each individual feature – first order statistics, formants, and harmonics – is able to achieve a prediction error less than the training data mean predictor in terms of MAE. The simple averaging of these predicted targets provides the best results in these tasks as well. While the proposed features and modeling are simple, we show that proposed approach is effective in various of speaker trait estimation tasks and outperform previously published results in these domains. To the best of authors knowledge, this is the first attempt to address the multilingual setting for speaker profiling tasks using short durations of speech data.

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 0.5 s. Also with around 1–2 s of data the MAE obtained is as good as the state-of-the-art results which were achieved using full duration of audio signal (> 10 s). This enables the system to be useful in speaker profiling, speaker recognition tasks, targeted advertisements in commercial applications with short audio recordings from the target speaker. The extension to noisy speech in conversational setting would be the next logical step to developing forensic speech applications.

Declaration of Competing Interest

The authors for the manuscript titled "Automatic speaker profiling from short duration speech data" declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Shareef Babu Kalluri: Conceptualization, Funding acquisition, Formal analysis, Writing - original draft, Writing - review & editing. **Deepu**

Vijayasenan: Conceptualization, Funding acquisition, Formal analysis, Writing - original draft, Writing - review & editing. **Sriram Ganapathy:** Conceptualization, Formal analysis, Writing - original draft, Writing - review & editing.

Acknowledgments

This work was partially funded by Science and Engineering Research Board (SERB) under grant no: EMR/2016/007934.

The authors would like to acknowledge the contribution of Sarthak Agrawal and Sanmathi Kamath who implemented the early version of the height/age prediction system while they were interning at the LEAP lab in Indian Institute of Science.

References

Arsikere, H., Leung, G.K., Lulich, S.M., Alwan, A., 2012. Automatic height estimation using the second subglottal resonance. In: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3989–3992.

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

Arsikere, H., Lulich, S.M., Alwan, A., 2011. Automatic estimation of the first subglottal resonance. J. Acoust. Soc. Am. 129 (5), EL197–EL203.

Arsikere, H., Lulich, S.M., Alwan, A., 2013b. Estimating speaker height and subglottal resonances using MFCCs and GMMs. IEEE Signal Process. Lett. 21 (2), 159–162.

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

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

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

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

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

Fitch, W.T., 1997. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. J. Acoust. Soc. Am. 102 (2), 1213–1222.

Fitch, W.T., Giedd, J., 1999. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. J. Acoust. Soc. Am. 106 (3), 1511–1522.

Ganchev, T., Mporas, I., Fakotakis, N., 2010. Audio features selection for automatic height estimation from speech. In: Proceedings of the 2010 Hellenic Conference on Artificial Intelligence. Springer, pp. 81–90.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., 1993. Darpa Timit Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1. NASA STI/Recon Technical Report N 93. Linguistic Data Consortium, Philadelphia.

Ghahremani, P., Nidadavolu, P.S., Chen, N., Villalba, J., Povey, D., Khudanpur, S., Dehak, N., 2018. End-to-end deep neural network age estimation. In: Proceedings of the 2018 Interspeech, pp. 277–281.

Gonzalez, J., 2003. Estimation of speakers' weight and height from speech: a re-analysis of data from multiple studies by lass and colleagues. Percept. Motor Skills 96 (1), 297–304.

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

Gonzalez, S., Brookes, M., 2014. Pefac-a pitch estimation algorithm robust to high levels of noise. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (2), 518–530.

Greisbach, R., 2007. Estimation of speaker height from formant frequencies. Int. J. Speech Lang. Law 6 (2), 265–277.

Hansen, J.H., Williams, K., Bořil, H., 2015. Speaker height estimation from speech: fusing spectral regression and statistical acoustic models. J. Acoust. Soc. Am. 138 (2), 1052–1067.

van Heerden, C., Barnard, E., Davel, M., van der Walt, C., van Dyk, E., Feld, M., Müller, C., 2010. Combining regression and classification methods for improving automatic speaker age recognition. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 5174–5177.

Jain, A.K., Ross, A., Prabhakar, S., et al., 2004. An introduction to biometric recognition. IEEE Trans. Circuits Syst. Video Technol. 14 (1), 4–20.

Kalluri, S.B., Vijayakumar, A., Vijayasenan, D., Singh, R., 2016. Estimating multiple physical parameters from speech data. In: Proceedings of the 26th IEEE International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, pp. 1–5.

Lass, N.J., Brown, W.S., 1978. Correlational study of speakers heights, weights, body surface areas, and speaking fundamental frequencies. J. Acoust. Soc. Am. 63 (4), 1218–1220.

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

Layer, J., Trudgill, P., 1979. 1. Phonetic and linguistic markers in speech. In: Trudgill//Social Markers in Speech. CUP, Cambridge, pp. 1–32. (1)

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

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

Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B., Bauer, J.G., et al., 2007. Comparison of four approaches to age and gender recognition for telephone applications. In: Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, 4. IEEE, pp. IV–1089.

Mporas, I., Ganchev, T., 2009. Estimation of unknown speaker's height from speech. Int. J. Speech Technol. 12 (4), 149–160.

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

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

Necioglu, B.F., Clements, M.A., Barnwell, T.P., 2000. Unsupervised estimation of the human vocal tract length over sentence level utterances. In: Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), 3. IEEE, pp. 1319–1322.

Nolan, F., 2005. Forensic speaker identification and the phonetic description of voice quality. In: A Figure of Speech: A Festschrift for John Laver. Psychology Press, pp. 385–411.

Pellom, B.L., Hansen, J.H., 1997. Voice analysis in adverse conditions: the centennial olympic park bombing 911 call. In: Proceedings of 40th Midwest Symposium on Circuits and Systems. Dedicated to the Memory of Professor Mac Van Valkenburg, 2. IEEE, pp. 873–876.

Peskin, B., Navratil, J., Abramson, J., Jones, D., Klusacek, D., Reynolds, D.A., Xiang, B., 2003. Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02. In: Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'03, 4. IEEE, pp. IV–792.

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

Poorjam, A.H., Bahari, M.H., Vasilakakis, V., et al., 2015. Height estimation from speech signals using i-vectors and least-squares support vector regression. In: Proceedings of the 38th International Conference on Telecommunications and Signal Processing (TSP). IEEE, pp. 1–5.

Poorjam, A.H., Bahari, M.H., et al., 2014. Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. In: Proceedings of the 4th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE, pp. 7–12.

Puts, D.A., Apicella, C.L., Cárdenas, R.A., 2012. Masculine voices signal men's threat potential in forager and industrial societies. Proc. R. Soc. B: Biol. Sci. 279 (1728), 601–609.

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

Rendall, D., Kollias, S., Ney, C., Lloyd, P., 2005. Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: the role of vocalizer body size and voice-acoustic allometry. J. Acoust. Soc. Am. 117 (2), 944–955.

Reynolds, D., 2002. An overview of automatic speaker recognition. In: Proceedings of the 2002 International Conference on Acoustics, Speech and Signal Processing (ICASSP) (S. 4072-4075).

Sadjadi, S.O., Ganapathy, S., Pelecanos, J.W., 2016. Speaker age estimation on conversational telephone speech using senone posterior based i-vectors. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5040–5044.

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

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

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., Narayanan, S., 2013. Paralinguistics in speech and language-state-of-the-art and the challenge. Comput. Speech Lang. 27 (1), 4–39.

Shivakumar, P.G., Li, M., Dhandhania, V., Narayanan, S.S., 2014. Simplified and supervised i-vector modeling for speaker age regression. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4833–4837.

Singh, R., Keshet, J., Hovy, E., 2016a. Profiling hoax callers. In: Proceedings of the 2016 IEEE Symposium on Technologies for Homeland Security (HST). IEEE, pp. 1–6.

Singh, R., Raj, B., Baker, J., 2016b. Short-term analysis for estimating physical parameters of speakers. In: Proceedings of the 4th International Conference on Biometrics and Forensics (IWBF). IEEE, pp. 1–6.

Smola, A., Vapnik, V., 1997. Support vector regression machines. Adv. Neural Inf. Process. Syst. 9, 155–161.

Souza, L.B.R.D., Santos, M.M.D., 2018. Body mass index and acoustic voice parameters: is there a relationship? Braz. J. Otorhinolaryngol. 84 (4), 410–415.

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

Tan, Z.-H., Lindberg, B., 2010. Low-complexity variable frame rate analysis for speech recognition and voice activity detection. IEEE J. Sel. Top. Signal Process. 4 (5), 798–807.

Tanner, D.C., Tanner, M.E., 2004. Forensic Aspects of Speech Patterns: Voice Prints, Speaker Profiling, Lie and Intoxication Detection. Lawyers & Judges Publishing Company.
Van Dommelen, W.A., Moxness, B.H., 1995. Acoustic parameters in speaker height and weight identification: sex-specific behaviour. Lang. Speech 38 (3), 267–287.
Walker, K., Strassel, S., 2012. The rats radio traffic collection system.. In: Proceedings of the 2012 Odyssey, pp. 291–297.
Williams, K.A., Hansen, J.H., 2013. Speaker height estimation combining GMM and linear regression subsystems. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 7552–7556.
Zazo, R., Nidadavolu, P.S., Chen, N., Gonzalez-Rodriguez, J., Dehak, N., 2018. Age estimation in short speech utterances based on LSTM recurrent neural networks. IEEE Access 6, 22524–22530.