# ESTIMATING MULTIPLE PHYSICAL PARAMETERS FROM SPEECH DATA

*Shareef Babu Kalluri, Ashwin Vijayakumar*,*
*Deepu Vijayasenan*

Department of E & C Engineering
National Institute of Technology Karnataka,
Srinivasnagar, Surathkal
Mangalore ,India - 575025

*Rita Singh*

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA-15213

## ABSTRACT

In this work, we explore prediction of different physical parameters from speech data. We aim to predict shoulder size and waist size of people from speech data in addition to the conventional height and weight parameters. A data-set with this information is created from 207 volunteers. A bag of words representation based on log magnitude spectrum is used as features. A support vector regression predicts the physical parameters from the bag of the words representation. The system is able to achieve a root mean square error of 6.6 cm for height estimation, 2.6cm for shoulder size, 7.1cm for waist size and 8.9 kg for weight estimation. The results of height estimation is on par with state of the art results.

***Index Terms***— Physical parameters, Speech forensics, Height, Weight, Shoulder size, Waist size

## 1. INTRODUCTION

Speech data not only contains information about the lingual message being conveyed but also the characteristics of the speaker. While the former is typically used in Automatic Speech Recognition, the latter information is effectively used in speaker identification and speaker verification. This is based on the natural assumption that speech depends on the the speech production system i.e. the physical characteristics of the person. One of the most common physical characteristics that can be predicted from speech data is the height of the speaker. Height estimation from speech is based on the assumption that the attributes of speech depend on the *vocal tract length* – that is correlated to the height of the person [1].

**Previous Work.** Different kinds of feature representations have been proposed in the past to study and predict the physical characteristics of the speaker using speech data. Some of the popular ones include *i-vector framework* using Mel Frequency Cepstral Coefficients (MFCC) [2] and features based on the statistics of frames [3]. More recently, a novel method [4] that uses estimated sub-glottal resonance frequencies to estimate speaker height was proposed. From an inspection of previous related works, it is apparent that there is no consensus about the most *suitable* feature that can be used for this task. Similarly, a wide variety of function approximation methods such as linear regression, support vector regression, artificial neural networks are employed to predict the height value. These systems are able to achieve Root Mean Square Error (RMSE) values of $6cm$.

Similar to height, weight estimation methods in the past have followed a similar approach and achieve a Pearson Correlation Coefficient (PCC) value of $0.52$ between the actual and estimated weights [2]. In addition to these physical characteristics, age is another interesting trait that researchers are able to successfully predict(For example [5, 6]). Smoking habit also has a telling effect on speech production and has been studied with considerable success [2].

**Overview and Contributions.** The key contributions of our work are:

1. In this work, we extend speech based predictions to two more physical characteristics – *shoulder width* and *waist size*, apart from height and weight. To facilitate the study of these characteristics, we collect a *new* speech dataset consisting of speech samples from 207 speakers. Additionally, we have speech samples in both English and the native language of the speaker – leading to diverse speech samples in about 12 languages native to India.
2. We hypothesize that both shoulder width and waist size influence the physical parameters of the lungs and the abdominal region, thereby influencing the speech production system. In line with our assumptions, we are able to predict these physical characteristics with considerable accuracy along with obtaining accuracies comparable to the *state-of-the-art* on height and weight

* Currently at Dept. of ECE, Virginia Tech

Thanks to Prof. Bhiksha Raj for valuable suggestions and discussions.

978-1-5090-0746-2/16/$31.00 ©2016 IEEE

| **Physical Characteristic** | **minimum** | **maximum** | **mean** | **standard deviation** |
|---|---|---|---|---|
| Height (*cm*) | 147 | 188 | 167.99 | 8.45 |
| Shoulder width (*cm*) | 30 | 53 | 43.50 | 3.72 |
| Waist (*cm*) | 64 | 112 | 84.72 | 7.77 |
| Weight (*kg*) | 39 | 107 | 64.47 | 12.09 |

**Table 1**. Mean, standard deviation, minimum and maximum values of each of the parameters in the data-set

parameters.

To the best of our knowledge, this work is the first to investigate the relationship between speech and the physical characteristics – shoulder width and waist size. This work is of immense importance in forensic applications. For example, predicting physical traits of the speaker can be very useful in tracing anonymous hoax calls.

In the next section, we discuss the details of the dataset followed by a description of our prediction algorithm in Section 3 method. In Section 4, we explain our experimental setup and report results and in Section 5, we summarize our contributions and conclude.

## 2. DATA COLLECTION

To study the correlation between the physical characteristics such as waist size, shoulder width, height and weight, we collected a new data set. Apart from collecting these pertinent parameters, we also obtained speech samples from volunteers. The dataset consists of samples from over 207 speakers (including 162 male and 45 female speakers) in the age group of 18 – 35 years. The speech samples were recorded in common environments such as a class-room, living-room and conference-hall. The recordings were made using a simple head-phone microphone (Logitech H110 stereo headset) at a sampling rate of 16 kHz. All speakers contributed roughly 2 minutes of data in 3 sessions – each lasting 40 seconds. The speakers were asked to read out loud sentences sampled from newspapers of both English and native Indian languages (depending on the first language of the speaker). The set of speakers in the dataset are linguistically diverse, consisting of a total of 12 different native tongues. Height, shoulder size and waist size are measured in centimetres and weight is measured in kilograms. Statistics of the physical parameters present in the dataset are provided in Table 1.

## 3. APPROACH

In this section, we explain in detail the procedure to predict the physical characteristics using speech data. Using all the training instances, we extract Short Time Fourier Transform (STFT) features. All the frame-level features so extracted are now clustered using *k-means* to construct a code-book. This code-book is then used to obtain a Bag of Words representation (BoW) of the utterance by quantizing features from each frame to the nearest cluster-centroid. Finally, these BoW features are input to the Support Vector Regressor (SVR) to predict the desired physical quantities of the speaker. An overview of the method is shown in Figure 1.

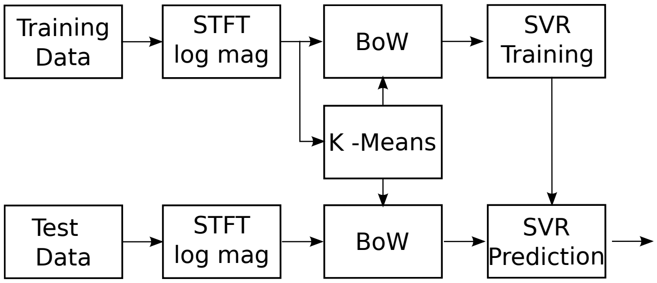


**Fig. 1**. Block diagram for speaker traits estimation

### 3.1. Feature Extraction

Various height/weight estimation methods in the past have explored a variety of features based such as Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC) [7], fundamental frequency and formants [8]. However, there is no consensus on the best feature while, most of the above features result in comparable performance. Therefore, in this study we choose to employ a BoW representation of the constituent frames. Each speech utterance is framed into 25 ms windows with 10 ms shift between successive frames. We evaluate using both the log-magnitude spectrum and the optional derivative features to form the BoW representation.

The extracted STFT features from the training samples is input to k-means clustering algorithm to obtain a code-book consisting of vectors, $\{\mathbf{m}_i\}, i{=}1, \ldots k$ – corresponding to the cluster centroids. The BoW representation for an utterance is obtained by counting the number of frames nearest to each vector in the code-book. Let $n_i$ be the number of speech frames nearest to the code-vector $\mathbf{m}$. Then, the BoW representation, $\mathbf{v}$ is obtained as $\frac{1}{N}[n_1, n_2, \ldots n_K]^T$, where $N$, the total number of frames is the normalization term.

### 3.2. Regression Model

Different linear and non-linear regression models have been experimented with in the context of physical parameter prediction [4, 7, 3]. In this work we use support vector regression [9] to predict the desired physical characteristics using the BoW features explained previously. Support vector regression is a general linear regression model that can be easily generalized to a non-linear regression model using a kernel.

Let the training set consist of $N_T$ instances, $\{\mathbf{v}_k, y_k\}$ where $k \in \{1, \ldots, N_T\}$, $\mathbf{v}_i$ is the input BoW representation and $y_i$, the physical characteristic to be predicted. Support vector regression aims to find a mapping $f(\cdot)$ that has a maximum deviation of $\epsilon$ from the target value. Thus errors are only penalized only if they are larger than $\epsilon$. Assuming a linear form for $f(\cdot)$ and with $L2$ regularization, the objective of the regressor is given by:

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

Here, $f(\mathbf{v}) = \mathbf{w}^T\mathbf{v} + \mathbf{b}$, $\mathbf{w}$ and $\mathbf{b}$ are the weight and bias terms of the linear function respectively.

The optimization of the SVR objective function is carried out in terms of the dot products of the data points among themselves. Therefore, the linear SVR can easily be extended to a non-linear SVR through the *kernel trick*. Drawing from previous work to estimate height [10], we employ the normalized polynomial kernel. The entries of the gram matrix of the kernel for degree $n \in \mathbb{N}$ are given by:

$$K(\mathbf{v}, \mathbf{v}') = \frac{(\mathbf{v}^T\mathbf{v}')^n}{\sqrt{(\mathbf{v}^T\mathbf{v})^n(\mathbf{v}'^T\mathbf{v}')^n}} \tag{2}$$

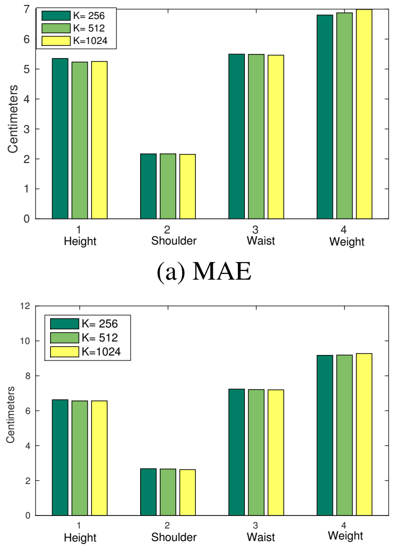
where $\mathbf{v}$ and $\mathbf{v}'$ are two data points – here, BoW features corresponding to speech utterances.

## 4. EXPERIMENTS

**Data.** We use the dataset described in Section 2 for the purpose of evaluation. The dataset into training and testing splits containing 137 speakers (104 male + 33 female) and 70 speakers ( 57 male + 13 female) respectively. Each of the recording sessions is considered as a different training/test samples and so, the training data contains 951 and 538 utterances respectively. The splits respect the linguistic diversity present in the whole set with proportional assignments of speakers. Each training/test utterance is converted to a 512-length BoW representation as described in Section 3.1.
**Metrics.** Following prior art, we report both RMSE and Mean Average Error (MAE) in all experiments.

**Analysis.** We perform an ablation study on the feature end by varying the dimension of the BoW representation and using the derivative features. We also investigate the variation of the error with the absolute value of the parameter. Finally, to demonstrate the scalability of our method to bigger datasets, we evaluate on the TIMIT [11] dataset for height-estimation (annotations for other physical parameters are not available). In the rest of this section, we explain the experimental setup and results in detail.

### 4.1. Feature Ablation Study



(a) MAE

(b) RMSE

**Fig. 2**. Comparison of the error metrics for all the physical parameters with the variation of $k$, the number of clusters. All values yield similar performance reflecting the robustness of our method with respect to $k$.

**log-magnitude STFT features** $(\log|STFT|)$**.** Initially, we only use the log-magnitude STFT features to obtain the BoW representation. The code-book used is constructed using different values of $k$, the number of clusters (256,512 and 1024). As shown in Figure 2, we obtain similar accuracy for all the vales proving that our method is generally robust to the number of clusters. For the rest of the results reported in this work, we use $k$=512

**Delta features** ($\delta$)**.** In order to add the dynamic information about the spectrum, delta and delta-delta features were added to the feature set. As before, a 512-dimensional BoW representation is extracted and input to the SVR to obtain predictions for the desired physical characteristics.

**Results.** Prediction results of this feature ablation study are provided in Table 2 for all the physical quantities. It can be noted that the RMSE is less than the standard deviation (see Table 1) of each of the physical parameter, thus making it a better predictor than just using the mean of the target values as the predictor. Moreover, the MAE and RMSE values of height estimation are similar to what is reported in previous work – RMSE of 6.2cm [4] and MAE of 5.3cm and RMSE of 6.8cm [3].

### 4.2. Analysis

In this section, we analyse the results reported in Table 2 and further study the source of error in detail.

As can be seen from Table 2, RMSE for height predictions does not improve significantly. However, there is a minor but steady improvement in other physical characteristics. In par-

| Metric | Feature | Height | Shoulder width | Waist size | Weight |
|---|---|---|---|---|---|
| MAE | $\log\|STFT\|$ | 5.23 | 2.17 | 5.49 | 6.88 |
| | $\log\|STFT\| + \delta$ | 5.20 | 2.12 | 5.40 | 6.72 |
| RMSE | $\log\|STFT\|$ | 6.56 | 2.66 | 7.21 | 9.19 |
| | $\log\|STFT\| + \delta$ | 6.58 | 2.57 | 7.08 | 8.91 |

**Table 2**. Comparision of both $\log|STFT|$ and full feature set with $\delta$ features. It can be seen that adding $\delta$ features improves the performance of the regressor indicating the importance of the dynamic content of the spectrum for this task.

ticular, the RMSE in weight reduces from 9.2 to 8.9. This reduction in error can be thought about as the reduction in the a-posterior variance of the parameter. We could thus infer that we are able to reduce the uncertainty in the parameter with the help of speech data for the physical characteristics of concern. We also achieve a reduction in the a-posterior variance by $2cm$ for height, $1cm$ for shoulder-width and by $3kg$ for weight estimation; while the waist size stays constant with negligible reduction in the error.

| Physical Characteristic | $\mathbf{p}<(\mu_\mathbf{p}-\sigma_\mathbf{p})$ | $(\mu_\mathbf{p}-\sigma_\mathbf{p})<$ $\mathbf{p}<(\mu_\mathbf{p}+\sigma_\mathbf{p})$ | $\mathbf{p}>(\mu_\mathbf{p}+\sigma_\mathbf{p})$ |
|---|---|---|---|
| Height | 4.05 | 5.62 | 9.88 |
| Shoulder | 1.81 | 2.49 | 3.79 |
| Waist | 11.57 | 4.01 | 12.38 |
| Weight | 3.99 | 7.36 | 15.77 |

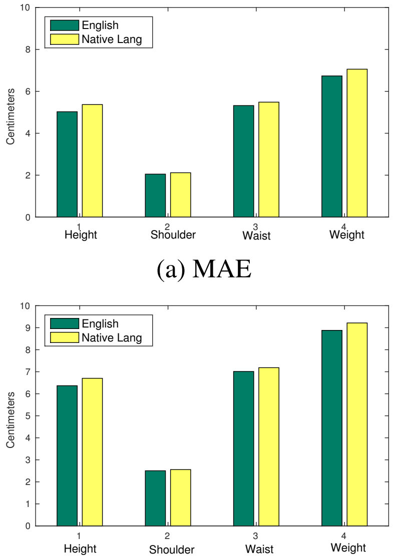
**Table 3**. RMSE values for the three sub groups of target values – low, mid, high.

To analyse the errors further we compute RMSE in different ranges of the absolute value of the desired physical characteristics. Let $\mu_p$,$\sigma_p$ are the mean and standard deviation of a parameter. We divide the test data into the following three sub-groups of the values of the parameter $p$ namely, (1) *low* – $(p<(\mu_p-\sigma_p))$, (2) *mid* – $((\mu_p-\sigma_p)<p<(\mu_p+\sigma_p))$ and (3) *high* – $(p>(\mu_p+\sigma_p))$. These divisions were performed *separately* for each physical characteristic. RMSE values of the predictions for each sub-group and each parameter is provided in Table 3. It can be observed that the errors are consistently small over all the subgroups for each characteristic except waist size. This could be because of the considerable error present in waist-size predictions for the complete test set. From Table 3, we can see that the error increases as we move up the range from *low* to *high* groups for each parameter value. This is acceptable as the relative error w.r.t the absolute value of the characteristic remains steady.

**Effect of Language.** We also analyse the effect of the language spoken on the prediction values to study the robustness of the method to language. To do this, we separate the training data into two consisting of only English and only native languages respectively. SVR models were trained on each set separately and the resulting performance is shown in Figure 3. It can be seen that both MAE and RMSE are better with only one language alone. The absolute difference in RMSE of estimation of height is better by $0.35$cm and the RMSE on weight is better by $0.3$kg. However, this may not imply that the method is not robust to language as there is comparatively less data for native languages. This bias is observed because of a skewed distribution of English-only speakers. Moreover, the number of native languages is relatively large (12 languages). Table 4 lists the performance using only English utterances. The performance improvement is of significance as it is obtained despite a decrease in the available training data while retaining the same range of the target values. (because of same speakers)

| Metric | Height | Shoulder width | Waist size | Weight |
|---|---|---|---|---|
| MAE | 5.02 | 2.05 | 5.32 | 6.73 |
| RMSE | 6.37 | 2.50 | 7.01 | 8.87 |

**Table 4**. MAE and RMSE values of predictions using only English language utterances of all speakers



(a) MAE

(b) RMSE

**Fig. 3**. Parameter prediction using utterances in English vs Native Languages

## 4.3. TIMIT evaluation

To demonstrate the scalability of our method, we evaluate on the TIMIT dataset [11] and predict height values (as the other characteristics are not available in the dataset). This dataset has more than 600 speakers across different age groups as compared to our data-set. However, the data-set has only height information. We obtain a MAE of $5.44cm$ and an

RMSE of $7.08cm$. The result is close to the *state-of-the-art* result of $6.8cm$ RMSE reported in Ganchev et.al [10].

## 5. SUMMARY AND CONCLUSION

In this work, we study the estimation of multiple physical characteristics of speakers using speech data. Apart from predicting height and weight as in previous works, we also predict shoulder-width and waist-size of the speaker. For this purpose, we collect a *new* dataset with 207 speakers. This dataset contains about 2 minutes of speech recordings in both English and the native language of the speaker.

We predict the physical characteristics using a BoW representation derived by clustering STFT based features of all the training frames. We use these features to train an SVR with a normalized polynomial kernal of degree 3. We obtain results close to the state of the art in height estimation – RMSE of 6.58 cm and MAE of 5.2 cm It can be observed that the a-posteriori variance of the weight is reduced to 9 Kg from 12 Kg using only speech data evidence. Similarly the shoulder size variance is reduced to 2.57 cm from 3.72 cm. The BoW representation seems to be consistent across predicting different physical parameters and across different languages – proving its robustness. Although restricting the task to only English utterances marginally improves the predictions. Overall this system achieves an RMSE of 6.37 cm for height estimation, 2.5cm for shoulder size, 7.01cm for waist size and 8.87 kg for weight estimation(See Table 4).

The current set of experiments are performed on a controlled environment for all the speakers. In addition each parameter estimation uses around 40 seconds of speech data. Overcoming these reductions is a promising direction for future efforts.

## 6. REFERENCES

[1] W Tecumseh Fitch and Jay Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

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

[3] Iosif Mporas Todor Ganchev and Nikos Fakotakis, "Automatic height estimation from speech in real-world setup," in *Proceedings of 18th European Signal Processing Conference (EUSIPCO)*, 2010.

[4] Harish Arsikere, Steven M Lulich, and Abeer Alwan, "Estimating speaker height and subglottal resonances using mfccs and gmms," *Signal Processing Letters, IEEE*, vol. 21, no. 2, pp. 159–162, 2014.

[5] Jitendra Ajmera and Felix Burkhardt, "Age and gender classification using modulation cepstrum.," in *Odyssey*, 2008, p. 25.

[6] Ming Li, Chi-Sang Jung, and Kyu Jeong Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition.," in *INTERSPEECH*, 2010, pp. 2826–2829.

[7] Sorin Dusan, "Estimation of speaker's height and vocal tract length from speech signal," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[8] Julio Gonzalez, "Estimation of speakers'weight and height from speech: A re-analysis of data from multiple studies by lass and colleagues," *Perceptual and motor skills*, vol. 96, no. 1, pp. 297–304, 2003.

[9] Alex J Smola and Bernhard Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[10] Todor Ganchev, Iosif Mporas, and Nikos Fakotakis, "Audio features selection for automatic height estimation from speech," in *Artificial Intelligence: Theories, Models and Applications*, pp. 81–90. Springer, 2010.

[11] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.