Estimation of Unknown Speaker's Height from Speech

Iosif Mporas[†] *and Todor Ganchev*[1,‡]

Wire Communications Laboratory, Dept. of Electrical and Computer Engineering,

University of Patras, 26500 Rion-Patras, Greece

[†]imporas@upatras.gr, [‡]tganchev@ieee.org

Abstract

In the present study, we propose a regression-based scheme for the direct estimation of the height of unknown speakers from their speech. In this scheme every speech input is decomposed via the *openSMILE* audio parameterization to a single feature vector that is fed to a regression model, which provides a direct estimation of the persons' height. The focus in this study is on the evaluation of the appropriateness of several linear and non-linear regression algorithms on the task of automatic height estimation from speech. The performance of the proposed scheme is evaluated on the TIMIT database, and the experimental results show an accuracy of 0.053 meters, in terms of mean absolute error, for the best performing Bagging regression algorithm. This accuracy corresponds to an averaged relative error of approximately 3%. We deem that the direct estimation of the height of unknown people from speech provides an important additional feature for improving the performance of various surveillance, profiling and access authorization applications.

Keywords: human height estimation from speech, speech processing, regression algorithms.

[1] Corresponding author: Todor Ganchev (tganchev@ieee.org)

1. Introduction

Over the last decades, plenty of effort has been invested in the exploration of biometric characteristics and the exploitation of purposely-developed technology for person authentication. This technology is nowadays used in applications such as access authorization, home arrest, forensic applications, call-centre related hidden authentications, which are transparent to the user, homeland security applications and anti-terror surveillance, branch-to-branch transactions, etc. Specifically, in these applications, among the most widely used biometric processes are face recognition, finger geometry recognition, fingerprint recognition, hand geometry recognition, iris and retina recognition, DNA analysis, and recognition of various voice-related biometric characteristics, such as speaker identity, etc.

Figure 1 summarizes the voice and non-voice biometric traits and the corresponding biometric processes which can be used in applications that require person authentication or in automatic surveillance applications. The biometrics processes that can be carried out through analysis of the speech of a user are grouped together under the label *speech processing-based trait analysis*.

_____

Figure 1

_____

Some of the biometric traits shown in the figure are referred to as soft biometric characteristics (skin colour, eye colour, build, weight, height, accent, etc) since they do not withstand the strict requirements of permanence and distinctiveness. Although their discriminative capacity does not permit the development of self-dependent biometric solutions they can be used as additional features for improving the robustness and accuracy of other biometric processes (Jain et al., 2004).

In particular, voice offers major advantages over the other types of biometric processes in terms of intrusiveness, cost, easy of deployment and user acceptance, since it is the least intrusive amongst the many biometrics that are being used. Apart from the recognition of the content of speech and the identity of the speaker, extra information such as accent, gender, age, body characteristics and other soft biometric traits of an individual can also be inferred from a speech utterance. Such information has particular importance in various practical applications, in situations such as automatic handling of 911-phone calls (or the equivalent 112-phone calls in Europe), where gender, age or language can be used to route the calls to the corresponding qualified personnel. Furthermore, such supplementary sources of information are also important for forensics and automated surveillance applications.

Due to the multifunctionality of speech, it offers the opportunity for extracting multiple layers of information from the speech waveform (Batliner and Huber, 2007; Cowie and Douglas-Cowie, 1995; Cowie et al., 2001). From the point of view of speech processing, speech conveys both explicit and implicit information. The explicit information is the linguistic content of the uttered message and the implicit information characterizes the speaker identity, attitude, emotional state, physiological traits, etc. While automatic speech recognition aims at capturing the explicit (linguistic) information of a spoken message (Junqua and Haton, 1995; Cole et al., 1998; Kuroiwa et al., 1999), voice biometrics focus on the decoding of the implicit (non-linguistic) biometric information which speech signals carry (Esposito et al., 2007). This non-linguistic information is related to the speaker's identity (Campbell, 1997; Beigi, 2010), gender (Zeng et al., 2006; Metze, 2007), age (Metze, 2007), dialect (Huang et al., 2007) and emotional state (Cowie et al., 2001), socio-economic status, personality, affective state and other aspects such as body size (González, 2006), etc. Besides, a number of studies have focused on the estimation of the physical characteristics of humans from speech, such as weight or height recognition.

Human height identification from speech has been studied over the last decades, both through subjective human evaluations (listeners' perception about height from speech) and approaches for automatic height estimation, while the corresponding conclusions have been particularly controversial (Gonzalez, 2003). Indeed, while in early studies (Lass and Davis, 1976; Lass and Brown, 1978; Lass et al., 1980; Lass et al., 1982) was found that listeners are able to estimate accurately speakers' heights from pre-recorded speech samples, in later studies opposite results were reported (Gunter and Manning, 1982; Kunzel, 1989). In recent research (van Dommelen and Moxness, 1995), listeners' estimations of height from speech were found significantly correlated with the actual heights only for the case of male speakers. Moreover, in (Gonzalez, 2003) the average accuracy on height estimations, made by human listeners, was found to not exceed 14%. More sophisticated studies (van Dommelen, 1993; Gonzalez, 2003), showed that human listeners are not able to accurately identify the heights of speakers, since they make their judgments according to vocal stereotypes regarding the speaker's body size.

Several studies have focused on the examination of correlations between the speaker's height and specific acoustic features, such as the fundamental frequency, the energy and the formants. Based on X-ray and magnetic resonance imaging, it was recently shown that the vocal tract length and the height of the speaker are strongly correlated in monkeys (Fitch, 1997) and humans (Fitch and Giedd, 1999). Although the speech production theory assumes correlation between the vocal tract length and the

formant frequencies (Fant, 1960), the formants have proved to not be correlated with the speaker's height (van Dommelen and Moxness, 1995; Gonzalez, 2003; Collins, 2000). The only exception was found in (Rendall et al., 2005), where it was shown that the forth formant and the height of male speakers are somehow correlated for the vowel schwa (van Oostendorp, 1998). It was also found that the fundamental frequency is not significantly correlated with the speakers' heights (Lass and Brown, 1978; Kunzel, 1989; Gonzalez, 2003; van Dommelen and Moxness, 1995). Furthermore, in (van Dommelen and Moxness, 1995) it was shown that the energy below 1 kHz is not correlated with the speaker's height. Finally, in (Dusan, 2005; Pellom and Hansen, 1997) the correlation of the Mel frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) and the linear prediction coefficients (LPC) (Makhoul, 1975) was examined and the highest correlation was observed for the seventh MFCC.

Various computational approaches for the estimation of human height from speech were reported in the literature. In (Pellom and Hansen, 1997), where a classification scheme was followed, the height scale was clustered to eleven classes of ±0.025 meters range, using Gaussian mixture models, and each audio file was assigned to one of the eleven clusters. In (Smith and Nelson, 2004) a scale factor, based on the EM algorithm (Hogg, 2005) and the formant frequencies, is correlated with the speakers' height. In (Blomberg and Elenius, 2009) speaker characteristics, among which the body height, are correlated with a search tree warp factor. In (Dusan, 2005) multiple linear regressions, on phone level, were utilized to estimate the speaker's height. Related work for vocal tract length and velum height estimation was presented in (Necioglu et al., 2000) and (Richmond, 1999), respectively.

In the present work, the height estimation task is considered as a numerical prediction problem and we examine the possibility of an automatic estimation of the height of an unknown speaker from his/her voice. In previous related work, the research efforts were restricted mostly to correlation measurements and investigation of the appropriateness of various acoustic parameters (Dusan, 2005; Smith and Nelson, 2004), or the height estimation task was reduced to a classification problem (Pellom and Hansen, 1997) by quantizing the height scale in a number of predefined intervals. In contrast, in the present study we propose a regression-based scheme for the direct estimation of the height of unknown speakers from speech. In the proposed scheme, every audio input is decomposed to a single feature vector, which serves as input to a regression model. The feature vector is based on the *openSMILE* audio parameterization and consists of a large set of statistical parameters, which are computed over a number of traditional frame-based acoustic parameters. Furthermore, we assess the importance of the attributes of the feature vector

with respect to the height estimation task, and identify the subsets of acoustic parameters that are beneficial in terms of height estimation accuracy. This is in contrast to previous related studies, where the feature vectors were simply the average values of the speech parameters, as in (Pellom and Hansen, 1997; Dusan, 2005). The main focus in the present study is on the evaluation of the appropriateness of several linear and non-linear regression algorithms on the task of automatic height estimation from speech. We evaluate the practical usefulness of the proposed regression scheme on the TIMIT database.

The remainder of this paper is organized as follows. In Section 2, we present in detail the regression-based scheme for the estimation of the height of unknown speakers. Section 3 offers a description of the experimental setup, the speech data used and the experimental procedure. In Section 4, we report the experimental results. Finally, in Section 5 we conclude this work.

2. Height Estimation from Speech using Regression

In the present study, we approach the height estimation task as a numerical prediction problem and not as a classification problem on a quantized space. Thus, here the target space, i.e. the speaker's height, h, is considered to be continuous, $h \in (\mathrm{H}_{\min}, \mathrm{H}_{\max})$, with $\mathrm{H}_{\min} > 0$ and maximum (biologically plausible) height $\mathrm{H}_{\max} << +\infty$, and subsequently the height estimation procedure is based on regression analysis.

In Fig. 2, we present the block diagram of the proposed regression-based scheme for the estimation of an unknown speaker's height from her/his voice sample (i.e. audio input). As can be seen in the figure, an unknown speaker pronounces a short utterance, which is typically of few seconds, that is being captured via a microphone, and next serves as input for the speech pre-processing and parameterization stages. Initially, the input audio signal is passed through a voice activity detector (VAD), and the non-speech intervals are excluded from the further processing. The remaining speech segments are passed through a speech parameterization engine, where frame-based acoustic features are computed. Afterwards, functional statistical parameters are computed over the frame-based acoustic features and a single feature vector V is formed. Finally, the feature vector V is utilized as input to a height estimation function $f(.)$. The height estimator utilizes a pre-trained regression model to produce the output $h=f(V)$, which is the estimated height of the unknown speaker.

Figure 2

We can formalize the height estimation scheme described above as follows. Let us define an audio input O originating from an unknown speaker. The audio input is split to N overlapping frames of length W, i.e. $O = o_1, o_2 ... o_N$. The VAD is applied to the input O and for each audio frame o_i, $1 \le i \le N$ a speech/non-speech decision is made, i.e. the observation-label pairs $\{o_i, l_i\}$ are defined as:

$$\{o_i, l_i\}, \text{ where } l_i = \begin{cases} 1, & when \quad o_i = speech \\ 0, & when \quad o_i = non\text{-}speech \end{cases}. \tag{1}$$

Afterwards, the sequence of speech frames is used to compute one parametric vector, which represents the corresponding speech parts of the initial audio input, O. In detail, initially, for each audio frame, o_i, a number of acoustic features, v_i^m, are computed using the corresponding m acoustic parameterization techniques, T_m, with $1 \le m \le M$, i.e.

$$v_i^m = T_m(o_i), \text{ where } v_i^m \in \mathbb{R}^{k(m)} \tag{2}$$

where $k(m)$ is the number of computed acoustic features by the parameterization technique T_m. At the second step, a number of K statistical values are computed from the acoustic features v_i^m and their corresponding VAD labels l_i, i.e.

$$V(k) = V_k = S_k\left(\{v^1, v^2, ..., v^m, ..., v^M\}, l_i\right), \tag{3}$$

where $V \in \mathbb{R}^K$ and $S_k(.)$ is the k th statistical function. The feature vector V, which represents the audio input O, is further used as input to a regression algorithm. Specifically, with respect to a height regression model, H, the regression algorithm will estimate the height, *h*, of the speaker, i.e.

$$h = f(V, H), \text{ where } h \in \mathbb{R} \text{ and } h \in (\mathrm{H_{min}}, \mathrm{H_{max}}), \text{ with } \mathrm{H_{min}} > 0 \text{ and } \mathrm{H_{max}} << +\infty. \tag{4}$$

The regression model, H, will come out of a labelled training set with a-priori known heights of the speakers and the feature vectors obtained through the speech parameterization procedure described above.

In an effort to investigate the appropriateness of a number of linear and non-linear regression algorithms, which have successfully been used on different numerical prediction tasks, such as forecasting (Vislocky and Fritsch, 1995), phone duration prediction (Yamagishia, 2008) and phonetic segmentation (Mporas et al., 2010), in the present work, we evaluate their performance on the height

estimation task. For the purpose of comprehensiveness in the following subsections, we review the regression techniques of interest.

2.1. Linear Regression (LR)

In linear regression (LR) all feature parameters are weighted and summed, i.e. the height estimation function f takes the form

$$h = w_0 + \sum_{k=1}^{K} w_k V_k \tag{5}$$

The attribute weights w_k are computed by applying the least-squares criterion over the training data,

$$\arg\min_{w_k} \left\{ \sum_{k=1}^{D} \left(h_{real}(k) - w_0 - \sum_{j=1}^{K} w_j V_j(k) \right)^2 \right\}, \tag{6}$$

where $h_{real}(k)$ is the real height of the k th speaker, D is the size of the training data, and w_0 stands for the bias.

Instead of using all attributes, M5' decision trees (refer to Section 2.4) can be applied for feature selection (Wang and Witten, 1997). During feature selection, the attribute with the smallest standardized coefficient is iteratively removed until no improvement is observed in the error estimation. The error estimation is given by the Akaike information criterion (Akaike, 1974) as:

$$AIC = 2g + D(b)\left(\ln\left(\frac{2\pi R_S}{D(b)} \right) + 1 \right), \tag{7}$$

where g is the number of parameters in the statistic model and R_s is the residual sum of squares:

$$R_S = \sum_{i=1}^{D} (h_{real} - h)^2 . \tag{8}$$

Here R_S indicates the cumulative squared error with respect to the real speaker heights, and a smaller value of the AIC indicates for a better model.

2.2. Multi-layer Perceptron Neural Networks (MLP)

Neural networks (NNs) with three layers have been proved capable for numerical predictions (Chester, 1990), since neurons are isolated and region approximations can be adjusted independently to each other. In detail, the output z_j of the j th neuron in the hidden layer of a multilayer perceptron (MLP) NN is defined as:

$$z_j = f\left(\sum_{k=1}^{K} w_{jk}^{(1)} V_k + w_{j0}^{(1)}\right), \quad j = 1, 2, ..., J, \tag{9}$$

where $f(x) = \left(1 + e^{-x}\right)^{-1}$ is the sigmoid activation function, J is the total number of neurons in the hidden layer, and $w_{jk}^{(1)}$ and $w_{j0}^{(1)}$ are the weight and bias terms, respectively. In the present work, the output layer of the MLP NN consists of a single un-thresholded linear unit, and the network output, h, is defined as:

$$h = \sum_{j=1}^{J} w_j^{(2)} \, z_j + w_0^{(2)} \tag{10}$$

All weights are adjusted during the training through the back propagation algorithm.

2.3. Support Vector Regression (SVR)

For the non-linear case of support vector regression (SVR) the two most widely used training algorithms are the ε-SVR (Vapnik, 1998) and the v-SVR (Scholkopf et al., 2000). In the present, we utilize the v-SVR because of its ability to adjust automatically the ε insensitive cost parameter. Given the set of training data $\{V(i), h_{real}(i)\}$ for each speaker i, with $V(i) = [V_1(i), V_2(i), ..., V_k(i), ..., V_K(i)]^T$, a function ϕ maps the attributes to a higher dimensional space. The primal problem of v-SVR,

$$\arg\min_{\mathbf{w}, \varepsilon, \xi_i, \xi_i^*} \left\{ \frac{1}{2} \mathbf{w}^T \mathbf{w} + C\left(v\varepsilon + \frac{1}{k} \sum_{i=1}^{k} (\xi_i + \xi_i^*) \right) \right\}, \tag{11}$$

is subject to the following restrictions: $\left(\mathbf{w}^T \phi(x_i) + \beta\right) - h_{real}(i) \le \varepsilon + \xi_i$, $h_{real}(i) - \left(\mathbf{w}^T \phi(x_i) + \beta\right) \le \varepsilon + \xi_i^*$, $\xi_i, \xi_i^* \ge 0$, with $\mathbf{w} \in \Box^N$, $\beta \in \Box$, $i \in [0, \mathrm{N}]$ and $\varepsilon \ge 0$. Here, ξ_i and ξ_i^* are the slack variables for exceeding the target value more or less than ε, respectively, and C is the penalty parameter. The kernel function is $K(\cdot,\cdot) = \phi(x)^T \phi(x)$. The value of v affects the number of support vectors and training errors. In the present work we rely on the radial basis kernel function $K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right)$.

2.4. Model Trees (M5')

In addition, we consider the M5' model tree algorithm proposed by (Wang and Witten, 1997), which is a rational reconstruction of M5 method developed by Quilan (1992). In tree structures, leaves represent classifications and branches represent conjunctions of attributes. The M5' tree is a binary decision tree constructed in two steps, namely the splitting and the pruning phase. During splitting, for each node the

algorithm computes the best attribute to split the T subset of data that reaches the node. The error criterion is the standard deviation of each class value that reaches each node. The attribute i with the maximum standard deviation reduction $\hat{\sigma}$ is selected for splitting that node, i.e.

$$\arg\max_i \left\{ \hat{\sigma} = \sigma(T) - \sum_j \frac{|T_{ij}|}{|T|} \times \sigma(T_{ij}) \right\}, \tag{12}$$

where T_{ij} are the subsets that result from splitting the node according to the chosen attribute i, with $1 \le i \le N$. The splitting process, which results to child nodes with smaller standard deviation, terminates when class values of the instances that reach a node have standard deviation equal to a small fraction of the original instance set, or if only few instances remain. When splitting is completed, a large tree structure will be constructed. For each node one linear regression model is calculated and simplified by dropping the attributes that do not reduce the expected error. The error for each node is the averaged difference between the predicted and the actual value of each instance of the training set that reaches the node. The computed error is weighted by the factor $(n+\nu)/(n-\nu)$, where n is the number of instances that reach that node and ν is the number of parameters in the linear model that give the class value at that node. This process is repeated until all the examples are covered by one or more rules. During the pruning phase, sub-trees are pruned if the estimated error for the linear model at the root of a sub-tree is smaller or equal to the expected error for the sub-tree.

2.5. Additive Regression (AR)

The Additive Regression (AR) (Friedman, 2002) is a meta-classifier (Witten and Frank, 2005) that enhances the performance of a regression base classifier. In each iteration, the algorithm fits a model to the residuals left by the classifier on the previous iteration. Prediction is accomplished by adding the predictions of each classifier. Reducing the shrinkage (learning rate) parameter helps prevent over-fitting and has a smoothing effect but increases the learning time.

Here, for regression base classifier we utilized the Decision Stump tree. Decision Stump algorithm performs regression (based on mean-squared error) or classification (based on entropy), while missing is treated as a separate value.

2.6. Bagging (Bag)

Furthermore, we rely on the Bagging algorithm (Breiman, 1996) to predict the height of unknown speakers. This method is a bootstrap ensemble of methods (bootstrap aggregation) that creates individual regression models by training the same learning algorithm on a random redistribution of the training set. Each regression model's training set is generated by randomly drawing, with replacements, N instances, where N is the size of the original training set. Many of the original instances may be repeated in the resulting training set while other may be left out. After the construction of several regression models, taking the average value of the prediction of each regression model gives the final description.

In this case, we applied the fast decision tree (Witten and Frank, 2005) as base learner. This learner builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back fitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces.

3. Experimental Setup

In the following, we describe the speech database used in the present evaluation, the feature vector that was computed for each audio file and the experimental procedure.

3.1. Speech Database

In this study, we utilized the TIMIT database (Garofolo, 1988). TIMIT is an American-English database containing microphone quality recordings of prompted speech. The speech waveforms are sampled with sampling frequency 16 kHz and resolution of 16 bits per sample.

TIMIT is provided with a standard division to Train and Test subsets. The Train subset consists of 462 speaker recordings (326 male and 136 female) and Test subset of 168 speaker recordings (112 male and 56 female). In both subsets, each speaker utters 10 sentences. All speakers cover seven major dialectal regions of the United States of America plus a set of speakers moving around.

The speakers' heights are provided with the database at the resolution of one inch. In the present study we follow the International System of Units (SI), i.e. the height is converted to meters, which results to the range [1.448, 2.032] meters, with mean and standard deviation values equal to 1.755 and 0.095 meters, respectively. The distribution of the TIMIT speakers' heights is shown in Figure 3. As can be seen in Figure 3, the heights range for the speakers from the Train subset is [1.448, 1.981] meters,

while for the Test subset this range is [1.524, 2.032] meters. In our experiments, the test speaker designated as MCTW0, with reported height of 2.032 meters, was excluded from the Test subset so that the heights of all speakers from the Test subset fit in the range of heights in the Train subset. The same procedure was followed in (Dusan, 2005; Pellom and Hansen, 1997).

Figure 3

3.2. Acoustic Features and Feature Vector

In the present study, we relied on the *openSMILE* acoustic parameterization tool (Eyben et al., 2009). In brief, initially, an energy-based voice activity detector (Young et al., 2006) is applied to each audio file, and the non-speech portions of the audio input are excluded from further processing. Next, for the speech portions, a number of acoustic parameters are computed on a frame-by-frame basis: (i) the zero-crossing rate (ZCR) from the time signal, (ii) the root mean square (RMS) frame energy, (iii) the fundamental frequency of speech normalized to 500 Hz, (iv) the harmonics-to-noise ratio (HNR) by autocorrelation function, and (v) the 12 first Mel frequency cepstral coefficients (MFCC), computed as in the HTK setup (Young et al., 2006). Here we made use of speech frame size of 20 milliseconds and the subsequent frames were overlapping by 10 milliseconds. Furthermore, the first temporal derivatives of the sixteen acoustic parameters, described to this end, were computed. Appended together, the 16 static parameters and their 16 temporal derivatives resulted to 32 acoustic features per speech frame. Finally, twelve functional statistics were computed over the 32 frame-level acoustic features obtained to this end. These are: (i) the mean, (ii) the standard deviation, (iii) the kurtosis, (iv) the skewness, (v-viii) the extreme values (minimum, maximum, relative position and range), and (ix-xii) the linear regression coefficients (offset, slope and their mean square error). Thus, the final feature vector consisted of 384 statistical parameters per audio file: 12 statistical functions applied for each of the 32 frame-level acoustic parameters. Therefore, the final feature vector did not include any of the frame-level acoustic parameters, but only the global file-level parameters that resulted from computing the 12 statistical functions over the 32 feature-level parameters. Further details about the *openSMILE* acoustic features are available in (Schuller et al., 2009).

3.3. Experimental Procedure

In all experiments, we followed a common experimental protocol. Specifically, we relied on the suggested Train/Test subset division as specified in the TIMIT database documentation. The Train subset was utilized for the training of the regression models, while the accuracy of automatic height estimation was measured on the Test subset of TIMIT. This resulted to 4620 and 1670 speech files (10 sentences per speaker) in the Train and Test subsets, respectively. In the present study, we consider only the case of height estimation for unknown speakers, i.e. there is no overlapping between the speakers in the Train and Test datasets.

For the training and evaluation of the regression algorithms described in Section 2, we utilized their Weka (Witten and Frank, 2005) and LibSVM (Chang and Lin, 2002) implementations. The parameter values of all regression models were empirically determined utilizing a bootstrap training set of 924 speech files, consisting of two sentences taken from each speaker of the Train subset.

4. Experimental Results

In order to evaluate the practical value of the proposed regression-based scheme for height estimation from speech, we performed experiments following the experimental protocol outlined in Section 3.3. In studies of this type, measures such as the square root of the average squared difference or simply the mean absolute difference are more appropriate statistics (Gonzalez, 2003). For this reason, in the present evaluation we adopt these two figures of merit, i.e. the mean absolute error (MAE) and the root mean squared error (RMSE), as indicators of the performance.

Previous research on the task of height estimation from speech has shown dependency of the problem to the gender of the speaker (van Dommelen and Moxness, 1995; Lass and Brown, 1978; Pellom and Hansen, 1997; Smith and Nelson, 2004). Particularly, in (Lass and Davis, 1976; Lass and Brown, 1978) it was observed that formant frequencies and height correlate oppositely for women than for men, which may be explained by the hormonally induced divergence in the laryngeal growth in males (Smith and Nelson, 2004; Pressman and Keleman, 1970). Thus, we present evaluation results for the gender-independent as well as for gender-dependent case.

As a first step, we evaluated the performance of the regression algorithms described in Section 2 on the Test set of TIMIT database. The experimental results are shown in Table 1, in terms of RMSE and MAE in meters. The best performing regression algorithm for each gender set are indicated in bold.

Table 1

As can be seen in Table 1, the best performance in terms of MAE was achieved by the bagging algorithm (*Bag*), for all the examined test sets, with respect to the gender of the speaker. For the gender-independent case the *LR* and *M5'* algorithms also achieved good performance of approximately 0.054 meters in terms of MAE. For the case of male speakers, the best performing *Bag* method achieved MAE of approximately 0.053 meters, followed by the *MLP* algorithm, which achieved MAE of approximately 0.056 meters. In the case of female speakers, the *Bag*, *MLP*, *M5'* and *SVR* algorithms achieved MAE of approximately 0.052 meters, and the lowest RMSE, equal to 0.064 meters, was achieved by the *Bag* and the *MLP* regression models.

In most of the cases, the RMSE was found to be higher but close enough to the values of MAE, which shows that few outlier predictions existed in the height estimations, and thus most of the regression-based height estimators showed robustness to gross errors. The error in the estimation of the heights of female speakers was found to be slightly lower than the male ones and the gender-dependent estimation more accurate than the gender-independent case, which is in agreement with (Gonzalez, 2003; Pellom and Hansen, 1997; Smith and Nelson, 2004).

As a second step, we examined the performance of the regression algorithms utilizing different subsets of acoustic features. In order to select the appropriate subsets of acoustic features we performed ranking of the attributes of the feature vectors computed from the Train subset of TIMIT. The ranking was performed with the Relief algorithm (Robnik-Sikonja and Kononenko, 1997), which evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.

The Relief algorithm computes a vector W of the estimations of the qualities of all the attributes. The ranking position of each attribute is defined by its ranking score, i.e. the corresponding estimation of quality, $w \in \square$, which indicates the degree of importance of that attribute. These ranking scores were used to cluster the acoustic features into five clusters using the EM algorithm (Hogg et al., 2005). In detail, the ranking scores, $w \in \square$, were used to iteratively train five one-dimensional Gaussian distributions. Each distribution modelled one cluster. After the completion of the EM training each acoustic feature (attribute) was assigned to the cluster were the corresponding ranking score had the

maximum likelihood. This clustering procedure ensures that attributes with close ranking scores will be grouped together in the same cluster, since their importance is alike, and thus, the resulting clusters will correspond to meaningful subsets of acoustic features.

The clusters obtained with the abovementioned procedure consisted of 26, 77, 127, 114 and 40 acoustic features, referred to with respect to the order of increasing importance. The subsets of acoustic features belonging to each cluster are shown in the Appendix.

In Table 2, we present the MAE and RMSE results, in meters, for the three gender sets and for different combinations of the clusters used as feature vectors. For each gender set, the best-performing combination of regression algorithm and feature clusters is indicated in bold. For the purpose of direct comparison, the results for the full feature set, presented in Table 1, are duplicated here denoted as cluster {1, 2, 3, 4, 5}. The size of the feature vector for clusters {1, 2, 3, 4, 5}, {1, 2, 3, 4}, {1, 2, 3}, {1, 2} and {1}, was 384, 344, 230, 103 and 26 parameters, respectively.

Table 2

As can be seen in Table 2, the use of the entire feature vector is suboptimal in terms of accuracy and computational demands. Indeed, for the gender-independent results and the results for male speakers, the best performing algorithm (*Bag*) achieved its lowest MAE scores utilizing the acoustic features of the four most important clusters, instead of the entire feature vector. In the case of female speakers, the best performance was observed for the *SVR* regression models, when used with the acoustic features of the two most important clusters. The top-ranked subsets of acoustic features selected by the ranking algorithm, have a higher relevance and the use of these subsets contributes for better utilization of the available training data and thus to a better accuracy of height estimation. For instance, in the case of female speakers, the use of a feature vector of 103 parameters instead of entire set of 384 acoustic features was observed to reduce the MAE by approximately 0.002 meters.

5. Discussion and Conclusion

In the present work, we studied a regression-based scheme for automatic estimation of the height of unknown humans from their speech. In this scheme, we compute one acoustic feature vector for each speech utterance that consists of statistical parameters, computed over a set of traditional frame-level acoustic features, which are estimated for stationary portions of the audio signal. This feature vector is

next used as input to a regression model, which provides an estimation of the speaker's height. The proposed regression-based scheme was evaluated with various implementations of the regression algorithm in an experimental setup, which considers automatic height estimation for unknown speakers. We followed the suggested Train-Test subset division of TIMIT, and the experimental results for the best performing Bagging regression algorithm showed an accuracy of 0.053 meters, in terms of mean absolute error.

The observed MAE of 0.053 meters, when related to the average height of the speakers in the database, 1.775 meters, corresponds to average estimation error of approximately 3%. This result is quite promising with respect to the results presented in the work of Gonzalez, (2003), where the average accuracy on height estimations made by human listeners was found to not exceed 14%. Thus, it can be concluded that the automatic height estimation methods can offer an advantageous performance, when compared to the one achieved by human listeners. This is considered of significant importance for forensic applications, and for applications related to automatic surveillance, remote tele-services and call-centres. In addition, the automatic height estimation from speech provides additional clues towards developing an automatic means for profiling of unknown people, i.e. people which were not 'seen' before, which could be beneficial in forensic and surveillance applications.

Since the height estimation from speech is complementary with the height estimation from video, the two can be used together for achieving robust estimation of the person's height. Preliminary results for the height estimation from video through calibrated cameras, reported for a small set of ten persons with known heights (Kispál and Jeges, 2008), show promising results -- standard deviation of the error of approximately 0.031 meters was observed. However, this approach only works when the entire body of the person is in the receptive view of the camera and there are no occlusions. In cases when a given person is very close to the camera or is partially occluded by other objects the height estimation from speech can still be useful source of information, given the availability of a short speech utterance.

Together with clues obtained from other information sources (body structure, colour maps, etc), the height estimation from speech is useful supporting feature that can be used for the re-identification of unknown people, when they re-appear in the field of view of a camera. Such functionality is required in applications related to public security, situation-aware smart-home environments, etc.

Finally, we deem that the presented height estimation scheme has the potential to become an important part of access authorization systems, which are based on speaker verification and speaker

identification. Specifically, the automatic height estimation from speech, which does not require extra effort during the user authentication process and thus remains transparent to the user, could be employed for improving the robustness of present-day speaker verification systems to impostor trials. This can be achieved either by including the estimated height as an additional parameter in the feature vector, or as pre-processing or/and post-processing with respect to the user model. In speaker identification systems, the automatic height estimation can be used as extra feature for narrowing the search range, and thus for reducing the computational and memory demands during the speaker identification process.

Acknowledgements

This work was partially supported by the Prometheus project (FP7-ICT-214901) “Prediction and Interpretation of human behaviour based on probabilistic models and heterogeneous sensors”, co-funded by the European Commission under the Seventh’ Framework Programme.

Appendix: Clusters obtained after the EM clustering of the feature ranking results

Cluster 1
RMS_En{min}, MFCC[2,3,5,8,11]{min}
MFCC[2-4,6-12]{mean}
MFCC[1,6,9,11]{max}
MFCC[2]{std}, MFCC_d[5]{std}
MFCC[8]{skewness}
MFCC[12]{linregc2}
MFCC[5]{linregerrQ}, MFCC_d[5]{linregerrQ}

Cluster 2
MFCC[1,4,6,7,9,10,12]{min}, MFCC_d[1]{min}, ZCR{min}, F0_d{min}
MFCC[5]{mean}, MFCC_d[1,2,4,5,7,11,12]{mean}
MFCC[2-4,7,8,10,12]{max}, MFCC_d[2,5]{max}, ZCR{max}, ZCR_d{max}, F0{maxPos}
MFCC[5,6,10-12]{std}, MFCC_d[2,7,8]{std}, HNR{std}, F0{std}
MFCC[1,2,9]{range}, MFCC_d[5,8]{range}, ZCR{range}
MFCC[2,4,6,7,9,10-12]{skewness}, MFCC_d[12]{skewness}, F0{skewness}
MFCC[11]{kurtosis}, ZCR{kurtosis}, F0{kurtosis}, F0_d{kurtosis}
MFCC[6,8,9,11]{linregc2}, MFCC_d[1,4,6,11,12]{linregc2}, ZCR_d{linregc2}
MFCC[2,6,12]{linregerrQ}, MFCC_d[2]{linregerrQ}, HNR{linregerrQ}, F0{linregerrQ}

Cluster 3
MFCC_d[3,5,8]{min}
MFCC[1]{mean}, MFCC_d[3,6,8-10]{mean}, RMS_En_d{mean}, ZCR_d{mean}, HNR{mean}, F0{mean}
MFCC[5]{max}, MFCC_d[7,8,12]{max}, HNR{max}, HNR_d{max}, F0{max}, F0_d{max}
MFCC[3,5-8,10-12]{range}, MFCC_d[1,2,7,9]{range}, ZCR_d{range}, HNR{range}, HNR_d{range}, F0{range}, F0_d{range}
MFCC[4,7-9]{std}, MFCC_d[3,4,6,9-12]{std}, HNR_d{std}, F0_d{std}
MFCC_d[1,2,6,8]{skewness}, RMS_En{skewness}, RMS_En_d{skewness}, ZCR{skewness}, ZCR_d{skewness}
MFCC[2-4,6-10,12]{kurtosis}, MFCC_d[1,2,4,8,9,12]{kurtosis}, HNR{kurtosis}
MFCC[6-8,10,11]{linregc1}, MFCC_d[1-4,6,7,10-12]{linregc1}, ZCR_d{linregc1}
MFCC[2-5,7,10]{linregc2}, MFCC_d[2,5,7-10]{linregc2}
MFCC[4,7-11]{linregerrQ}, MFCC_d[3,4,6-10,12]{linregerrQ}, F0_d{linregerrQ}
MFCC[6,11]{minPos}, F0_d{minPos}
MFCC[2,9]{maxPos},MFCC_d[1,2,12]{maxPos}, HNR{maxPos}, F0_d{maxPos}

Cluster 4
MFCC_d[2,4,6,7,9-12]{min}, F0{min}, RMS_En_d{min}, ZCR_d{min}, HNR_d{min}
HNR_d{mean}, F0_d{mean}, RMS_En{mean}
MFCC_d[1,3,4,6,9-11]{max}, RMS_En{max}, RMS_En_d{max}
MFCC[1,3]{std}, MFCC_d[1]{std}, RMS_En{std}, RMS_En_d{std}, ZCR{std}
MFCC[4]{range}, MFCC_d[3,4,6,10-12]{range}, RMS_En{range}, RMS_En_d{range}
MFCC[1,5]{kurtosis}, MFCC_d[3,5-7,10,11]{kurtosis}, RMS_En{kurtosis}, RMS_En_d{kurtosis}, ZCR_d{kurtosis}, HNR_d{kurtosis}
MFCC[1,5]{skewness}, MFCC_d[3-5,7,9-11]{skewness}, HNR{skewness}, HNR_d{skewness}, F0_d{skewness}
MFCC[1-5,9,12]{linregc1}, MFCC_d[5,8,9]{linregc1}, RMS_En{linregc1}, RMS_En_d{linregc1}, ZCR{linregc1}, HNR{linregc1}, HNR_d{linregc1}, F0{linregc1}, F0_d{linregc1}
MFCC[1,3]{linregc2}, HNR_d{linregc2}, RMS_En_d{linregc2}, RMS_En{linregc2}, HNR{linregc2}, F0{linregc2}, F0_d{linregc2},
MFCC[1,3]{linregerrQ}, MFCC_d[1,11]{linregerrQ}, RMS_En{linregerrQ}, RMS_En_d{linregerrQ}, HNR_d{linregerrQ}
MFCC[8,10,12]{maxPos}, MFCC_d[4,6,8,10]{maxPos}, RMS_En{maxPos}, HNR_d{maxPos}
MFCC[5,7,8,12]{minPos}, MFCC_d[6,7,11,12]{minPos}, HNR_d{minPos}, F0{minPos}

Cluster 5
MFCC[1-4,9,10]{minPos}, MFCC_d[1-5,8-10]{minPos}, RMS_En{minPos}, RMS_En_d{minPos}, ZCR{minPos}, ZCR_d{minPos}, HNR{minPos}
MFCC[1,3-7,11]{maxPos}, MFCC_d[3,5,7,9,11]{maxPos}, RMS_En_d{maxPos}, ZCR{maxPos}, ZCR_d{maxPos}
HNR{min}
ZCR{mean}
ZCR_d{std}
ZCR{linregc2}
ZCR{linregerrQ}, ZCR_d{linregerrQ}

References

Akaike, H. (1974). A new look at the statistical model identification. In *IEEE Transactions on Automatic Control*, 19(6):716-723.

Batliner, A., Huber, R. (2007). Speaker characteristics and emotion classification. In Speaker Classification I, LNAI 4343, C. Műller (Ed.), Springer 2007, pp.138-151.

Beigi, H. (2010). Fundamentals of Speaker Recognition. ISBN-13: 978-0387775913, Springer, 2010.

Blomberg, M. and Elenius, D. (2009). Estimating speaker characteristics for speech recognition. In *Proc. of the XXIIth Swedish Phonetics Conference* (*FONETIK 2009*), pp. 154-158.

Breiman, L. (1996). Bagging predictors. In *Machine Learning*, 24(2):123-140.

Campbell, J. P. (1997). Speaker recognition: a tutorial. In *Proceedings of the IEEE*, 85(9), September 1997.

Chang, C. C. and Lin, C. J. (2002). Training v-support vector regression: theory and algorithms. In *Neural Computation*, 14(8):1959–1977.

Chester, D. L. (1990). Why two hidden layers are better than one. In *Proc. of the International Joint Conference on Neural Networks*. Vol. 1, pp. 265-268.

Cole et al. (1988). Survey of the State of the Art in Human Language Technology (Studies in Natural Language Processing). Editors: R. Cole, J, Mariani, H. Uszkoreit, G. Battista Varile, A. Zaenen, A. Zampolli, ISBN-13: 978-0521592772, Cambridge University Press.

Collins, S. A. (2000). Men's voices and women's choices. In *Animal Behaviour*, 60:773-780.

Cowie, R. and Douglas-Cowie, E. (1995). Speakers and hearers are people: Reflections on speech deterioration as a consequence of acquired deafness. In *Profound Deafness and Speech Communication*, K-E. Spens and G. Plant, Eds. London, UK: Whurr, pp. 510-527.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. In *IEEE Signal Processing Magazine*, 18(1):32-80.

Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Transactions Acoustics, Speech and Signal Processing*, 28(4):357–366.

van Dommelen, W. A. (1993). Speaker height and weight identification: re-evaluation of some old data. In *Journal of Phonetics*, 21:337-341.

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

Dusan, S. (2005). Estimation of speaker's height and vocal tract length from speech signal. In *Proc. of the 9th European Conference on Speech Communication and Technology* (*Interspeech 2005*), pp. 1989-1992.

Esposito, A., Bratanic, M., Keller, E. and Marinaro, M. (2007). Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue. Vol. 18 NATO Security through Science Series - E: Human and Societal Dynamics, IOS Press.

Eyben, F., Wöllmer, M. and Schuller, B. (2009). OpenEAR - introducing the Munich open-source emotion and affect recognition toolkit. In *Proc. of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction* (*ACII 2009*).

Fant, G. (1960). Acoustic Theory of Speech Production, Mouton, The Hague.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaque. In *Journal of Acoustical Society of America* (*JASA*), 102(2):1213-1222.

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

Friedman, J. H. (2002). Stochastic gradient boosting. In *Computational statistics and data analysis*, 38(4):367-378.

Garofolo, J. (1988). Getting Started with the DARPA-TIMIT CD-ROM: an acoustic phonetic continuous speech database. National Institute of Standards and Technology (NIST), Gaithersburgh, MD, USA.

Gonzalez, J. (2003). Estimation of speaker's weight and height from speech: a re-analysis of data from multiple studies by Lass and colleagues. In *Perceptual and Motor Skills*, 96:297-304.

González, J. (2006). Research in acoustics of human speech sounds: correlates and perception of speaker body size. Book chapter in S.G. Pandalai. *Recent Research Developments in Applied Physics*, vol. 9. Edited by Transworld Research Network, Trivadrum-695 023, Kerala. [ISBN: 81-7895-213-0]

Gunter, C. D. and Manning, W. H. (1982). Listener estimations of speaker height and weight in unfiltered and filtered conditions. In *Journal of Phonetics*, 10:251-257.

Hogg, R., McKean, J. and Craig, A. (2005). Introduction to Mathematical Statistics, Upper Saddle River, NJ: Pearson Prentice Hall, pp. 359-364.

Huang, R., Hansen, J. H. L. and Angkititrakul, P. (2007). Dialect/accent classification using unrestricted audio. In *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):453-464.

Jain, A. K., Dass, S. C. and Nandakumar, K. (2004). Can soft biometric traits assist user recognition? In *Biometric Technology for Human Identification*. Edited by Jain, A. K. and Ratha, N. K. *Proceedings of the SPIE 2004*, vol. 5404, pp. 561-572.

Junqua, J.-C. and Haton, J.-P. (1995). Robustness in Automatic Speech Recognition -- Fundamental and Applications, Kluwer Academic Publishers, ISBN-13: 978-0792396468.

Kispál I. and Jeges, E. (2008). Human height estimation using a calibrated camera. In *Proc. of the Computer Vision and Pattern Recognition* (*CVPR 2008*).

Kunzel, H. J. (1989). How well does average fundamental frequency correlate with speaker height and weight? In *Phonetica*, 46:117-125.

Kuroiwa, S., Naito, M., Yamamoto, S. and Higuchi, N. (1999). Robust speech detection method for telephone speech recognition system. In *Speech Communication*, 27:135-148.

Lass, N. J. and Davis, M. (1976). An investigation of speaker height and weight identification. In *Journal of Acoustical Society of America* (*JASA*), 60(3):700-703.

Lass, N. J. and Brown, W. S. (1978). Correlation study of speaker's heights, weights, body surface areas, and speaking fundamental frequencies. In *Journal of Acoustical Society of America* (*JASA*), 63(4):700-703.

Lass, N. J., Phillips, J. K. and Bruchey, C. A. (1980). The effect of filtered speech on speaker height and weight identification. In *Journal of Phonetics*, 8:91-100.

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

Makhoul, J. (1975). Linear prediction: a tutorial review. In *Proceedings of the IEEE*, 63(5):561-580, April 1975.

Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Müller, C., Huber, R., Andrassy, B., Bauer, J. and Littel, B. (2007). Comparison of four approaches to age and gender recognition for telephone applications. In *Proc. of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP 2007*). Vol. 4, pp. 1089-1092.

Mporas, I., Ganchev, T. and Fakotakis, N. (2010). Speech segmentation using regression fusion of boundary predictions. In *Computer Speech and Language*, 24(2):273-288.

Necioglu, B. F., Clements, M. A. and Barnwell III, T. P. (2000). Unsupervised estimation of the human vocal tract length over sentence level utterances. In *Proc. of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP 2000*). Vol. 3, pp. 1319-1322.

van Oostendorp, M. (1998). Schwa in phonological theory, In *GLOT International*, 3:3-8.

Pellom, B. L. and Hansen, J. H. L. (1997). Voice analysis in adverse conditions: the centennial Olympic park bombing 911 call. *In Proc. of the 40th Midwest Symposium on Circuits and Systems* (*MWSCAS 1997*). Vol. 2, pp. 873-876.

Pressman, J. J. and Keleman, G. (1970). Physiology of the Larynx (Rev. by J. A. Krichner). Rochester, Minnesota: American Academy of Ophthalmology and Otolaryngology, 1970.

Quilan, J. R. (1992). Learning with continuous classes. In *Proc. of the 5th Australian Joint Conference on Artificial Intelligence*, World Scientific, pp. 343-348.

Richmond, K. (1999). Estimating velum height from acoustics during continuous speech. In *Proc. of the 6th European Conference on Speech Communication and Technology* (*Eurospeech 1999*). Vol. 1, pp. 149-152.

Rendall, D., Kollias, S. and Ney, C. (2005). Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: the role of vocalizer body size and voice-acoustic allometry. In *Journal of Acoustical Society of America* (*JASA*), 117(2):1-12.

Robnik-Sikonja, M. and Kononenko, I. (1997). An adaptation of Relief for attribute estimation in regression. In *Proc. of the 14th International Conference on Machine Learning*, pp. 296-304.

Scholkopf, B., Smola, A., Williamson, R. and Bartlett, P. L. (2000). New support vector algorithms. In *Neural Computation*, 12(5):1207-1245.

Schuller, B., Steidl, S. and Batliner, A. (2009). The Interspeech 2009 emotion challenge. In *Proc. of the 10th Annual Conference of the International Speech Communication Association* (*Interspeech 2009*), pp. 312-315.

Smith, L. H. and Nelson, D. J. (2004). An estimate of physical scale from speech. In *Proc. of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP 2004*). Vol. 1, pp. 561-564.

Vapnik, V. (1998). Statistical Learning Theory, New York: Wiley.

Vislocky, R. L. and Fritsch, J. M. (1995). Generalized additive models versus linear regression in generating probabilistic MOS forecasts of aviation weather parameters. In *Weather and Forecasting*, 10(4):669–680.

Wang, Y. and Witten, I. H. (1997). Inducing model trees for continuous classes. In *Proc. of the 9th European Conference on Machine Learning*, pp. 128-137.

Witten, H. I. and Frank, E. (2005). Data Mining: practical machine learning tools and techniques. Morgan Kaufmann Publishing.

Yamagishia, J., Kawaia, H. and Kobayashib, T. (2008). Phone duration modeling using gradient tree boosting. In *Speech Communication*, 50(5):405-415.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. (2006). The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department.

Zeng, Y., Wu, Z., Falk, T. H. and Chan, W.-Y. (2006). Robust GMM-based gender classification using pitch and RASTA-PLP parameters of speech. In *Proc of Intl. Conf. on Machine Learning and Cybernetics*, 2006.

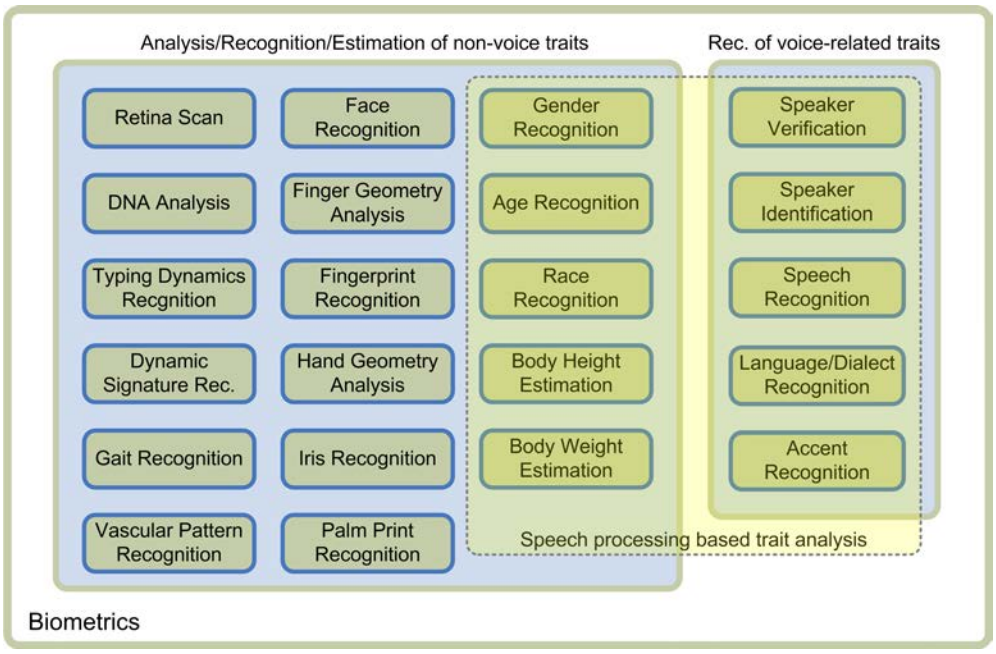

Figure 1. Commonly used biometric processes for analysis of voice and non-voice traits

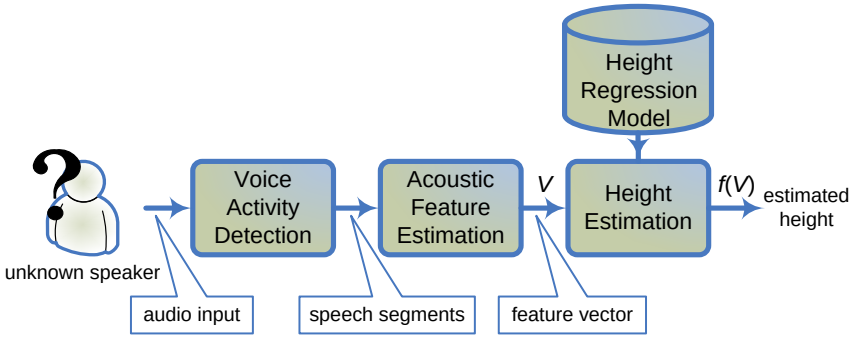

Figure 2. Block diagram of the regression-based automatic height estimation from speech.

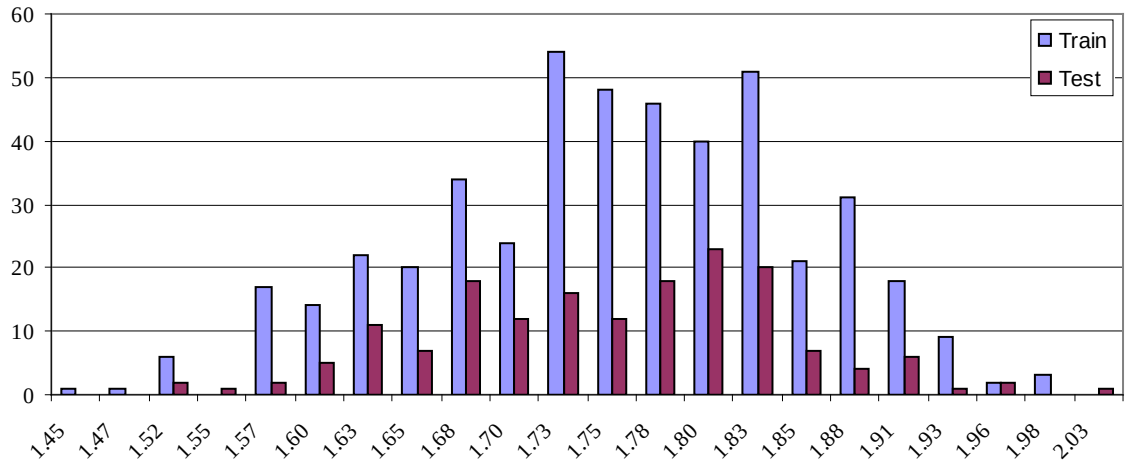

Figure 3. Histogram of the distribution of speakers in the Train and Test subsets of TIMIT with respect to their height. The abscissa stands for the speaker's height and the ordinate for the number of speakers for each height value.

Table 1. Height estimation in terms of MAE and RMSE (in meters).

Regression Method	All Speakers		Male Speakers		Female Speakers	
	MAE	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
AR	0.056	0.072	0.057	0.072	0.053	0.069
Bag	**0.053**	**0.067**	**0.053**	**0.068**	**0.052**	**0.064**
LR	0.054	0.069	0.065	0.082	0.056	0.073
M5'	0.054	0.069	0.059	0.076	**0.052**	0.068
MLP	0.087	0.104	0.056	0.073	**0.052**	**0.064**
SVR	0.054	0.069	0.057	0.072	**0.052**	0.068

Table 2. Height estimation in terms of MAE and RMSE (in meters) for different subsets of acoustic features.

Feature Clusters	Regression Method	All Speakers		Male Speakers		Female Speakers	
		MAE	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
{1,2,3,4,5}	*AR*	0.056	0.072	0.057	0.072	0.053	0.069
{1,2,3,4,5}	*Bag*	**0.053**	**0.067**	**0.053**	0.068	0.052	**0.064**
{1,2,3,4,5}	*LR*	0.054	0.069	0.065	0.082	0.056	0.073
{1,2,3,4,5}	*M5'*	0.054	0.069	0.059	0.076	0.052	0.068
{1,2,3,4,5}	*MLP*	0.087	0.104	0.056	0.073	0.052	**0.064**
{1,2,3,4,5}	*SVR*	0.054	0.069	0.057	0.072	0.052	0.068
{1,2,3,4}	*AR*	0.056	0.072	0.057	0.072	0.053	0.069
{1,2,3,4}	*Bag*	**0.053**	**0.067**	**0.053**	**0.067**	0.052	**0.064**
{1,2,3,4}	*LR*	0.054	0.069	0.065	0.081	0.056	0.073
{1,2,3,4}	*M5'*	0.055	0.071	0.060	0.075	0.057	0.074
{1,2,3,4}	*MLP*	0.087	0.104	0.056	0.073	0.105	0.131
{1,2,3,4}	*SVR*	**0.053**	0.069	0.056	0.071	0.051	0.066
{1,2,3}	*AR*	0.056	0.072	0.057	0.072	0.055	0.069
{1,2,3}	*Bag*	0.054	0.068	**0.053**	0.068	0.051	0.065
{1,2,3}	*LR*	**0.053**	0.068	0.061	0.077	0.054	0.070
{1,2,3}	*M5'*	**0.053**	0.068	0.061	0.078	0.057	0.073
{1,2,3}	*MLP*	0.099	0.136	0.111	0.141	0.119	0.146
{1,2,3}	*SVR*	0.053	0.068	0.055	0.070	0.051	0.066
{1,2}	*AR*	0.055	0.071	0.059	0.074	0.054	0.068
{1,2}	*Bag*	0.054	0.069	**0.053**	0.068	0.051	**0.064**
{1,2}	*LR*	**0.053**	0.068	0.058	0.073	0.052	0.067
{1,2}	*M5'*	0.054	0.069	0.059	0.076	0.051	0.066
{1,2}	*MLP*	0.119	0.157	0.100	0.129	0.089	0.114
{1,2}	*SVR*	**0.053**	0.068	0.055	0.070	**0.050**	0.065
{1}	*AR*	0.059	0.074	0.058	0.072	0.054	0.068
{1}	*Bag*	0.056	0.070	0.054	0.069	0.054	0.066
{1}	*LR*	0.055	0.070	0.055	0.070	0.052	0.066
{1}	*M5'*	0.056	0.071	0.058	0.074	0.055	0.069
{1}	*MLP*	0.074	0.145	0.093	0.132	0.108	0.215
{1}	*SVR*	0.055	0.069	0.054	0.069	0.053	0.067

List of Figures:

Figure 1. Commonly used voice and non-voice biometric processes for analysis of voice and non-voice traits. 2

Figure 2. Block diagram of the regression-based automatic height estimation from speech. 5

Figure 3. The histogram shows the distribution of speakers in the Train and Test subsets of TIMIT with respect to their height. The abscissa stands for the speaker's height and the ordinate for the number of speakers that reported that specific value. 11

List of Tables:

Table 1. Height estimation in terms of MAE and RMSE (in meters). 13

Table 2. Height estimation in terms of MAE and RMSE (in meters) for different subsets of acoustic features. 14

Clusters obtained after the EM clustering of the feature ranking results Appendix