AUTOMATIC HEIGHT ESTIMATION FROM SPEECH IN REAL-WORLD SETUP

Todor Ganchev, Iosif Mporas, and Nikos Fakotakis

Wire Communications Laboratory, Dept. of Electrical and Computer Engineering,
University of Patras, 26500 Rion-Patras, Greece
phone: +302610969808, fax: +302610997336, email: tganchev@ieee.org
web: http://www.wcl.ece.upatras.gr/ai

ABSTRACT

We propose a Gaussian process based regression scheme that provides a direct estimation of the height of unknown speakers and is applicable to real-world autonomous surveillance applications. This scheme relies on utterance-level speech parameterization followed by regression modelling, which estimates the height of the speaker and the uncertainty interval of that estimation. Experiments on the TIMIT database demonstrated that a feature vector composed of the top-50 ranked parameters offers a good trade-off between computational demands and accuracy. The proposed scheme for automatic height estimation was evaluated in the smart-home and public security scenarios offered by the PROMETHEUS database. The averaged relative error of height estimation remained approximately 3%, in both indoor and outdoor conditions, which indicates the good robustness of the proposed scheme.

1. INTRODUCTION

Nowadays, applications such as data/area access authorization, remote user authentication, forensic applications, homeland security applications and anti-terror surveillance, etc have become common. In these applications, among the most widely used biometric processes are fingerprint recognition, finger/hand geometry recognition, face recognition, iris and retina recognition, DNA analysis, and recognition of various voice-related biometric characteristics, such as the speaker voiceprint. Some other biometric traits, referred to as soft biometric characteristics are the skin colour, the eye colour, the body build, weight, and height and the accent. Although the discriminative capacity of the soft biometric characteristics does not allow the development of self-dependent biometric solutions, they were found useful as additional features for improving the robustness and accuracy of other biometric processes [1].

Given specific controlled conditions, human height can be estimated from images and video sequences. An early study on height estimation from video through calibrated cameras, reported for a small set of ten persons with known heights [2], has demonstrated promising results – a standard deviation of the estimation error of approximately 0.031 meters was observed. However, this approach works only when the entire body of the person is in the receptive view of the camera and no occlusions occur. A recent work [3] demonstrated height estimation accuracy of approximately 0.0267 meters for a set of 127 people from different ages and demographics. However, this approach assumes that the persons' face is oriented towards the camera, which limits the area of its applicability.

Related work on height detection from speech was reported in [4], where the height scale was split to eleven classes of ±0.025 meters range, and for each class a Gaussian mixture model was built. Afterwards, each input audio file was assigned to one of the eleven clusters. In [5] a scale factor, based on the EM algorithm and the formant frequencies, was correlated with the speakers' height. In [6] speaker characteristics, among which the body height, were correlated with a search tree warp factor. Multiple linear regression algorithms were utilized on phone level [7], and several non-linear regression algorithms were applied on utterance level [8] for estimating the speaker's height.

In the present work, we focus on human height estimation from speech in two real-world applications: smart-home and public security (bankomat i.e. ATM and airport surveillance). Both applications involve uncontrolled operational conditions, such as occlusions of the persons of interest by (moving) objects or other humans, humans partially or completely outside the perceptive view of the cameras, bended human bodies, various hats, open umbrellas, etc. In these cases, accurate height estimation from video is not always possible, and this complicates the person re-identification among different cameras or after her/his reappearing in the scene. In such cases, given the availability of a speech utterance, height estimation from speech can be indispensable.

The approach discussed here differs from previous related work in that:

(i) we do not rely on the basic speech features (formants, pitch, energy, MFCCs, etc), which in earlier studies were reported to be weakly or moderately correlated to height but instead use utterance-level audio parameters that are derived through statistical processing of frame-level speech features,

(ii) we systematically select the most useful attributes among the 6552 statistical parameters offered by the openSMILE audio parameterization [9],

(iii) we perform a direct estimation of the speakers height via Gaussian process (GP) regression that also provides the uncertainty interval of each estimation, and

(iv) we aim at the estimation of the height of unknown speakers in real-word setup: smart-home and public security scenarios.

 ISSN 2076-1465

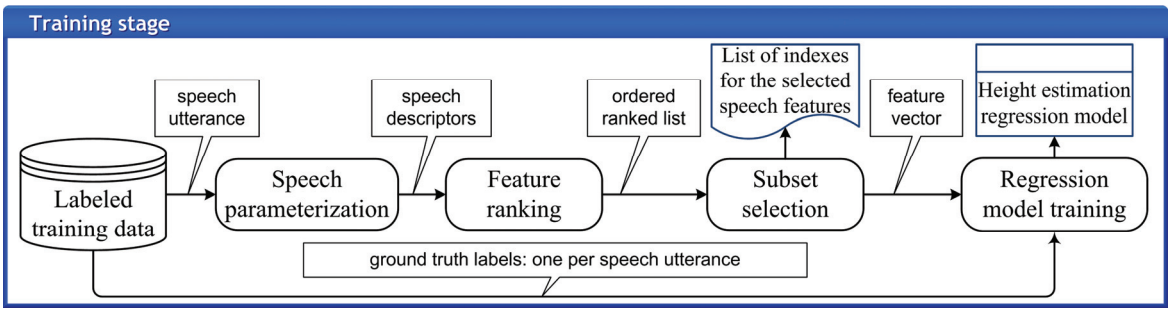


Figure 1 – Regression-based scheme for height estimation from speech for unknown speakers: the training stage.

The probabilistic nature of the proposed regression scheme facilitates fusion with the estimations from other sensors, including video cameras, which is quite important for the use of this technology in practical applications.

2. HUMAN HEIGHT ESTIMATION FROM SPEECH

The efforts for human height estimation from speech are based on the assumption that there is a strong correlation between the height of a person and the length of her/his vocal tract. Studies with X-ray and magnetic resonance imaging (MRI) provide evidence in support of that assumption [10]. Furthermore, the speech production theory [11] assumes that the vocal tract length and the formant frequencies of speech are correlated, and therefore human height can be inferred from speech. Although it was experimentally found that among the speech formants only the forth one is correlated with the human height, other speech descriptors (LPC, MFCC, etc) were reported to be correlated as well.

In the present work, we view the automatic human height estimation from speech as a supervised learning task that aims at the creation of regression model, $\mathrm{f}(\mathbf{x})$, from a given training set $\mathbf{D}=\{\mathbf{X},\mathbf{h}\}=\{\mathbf{x}_i^{(d)},h_i \mid i=1,2,...,n\}$. Here $\mathbf{D}$ consists of n d-dimensional feature vectors $\mathbf{x}_i^{(d)}$, which are computed from the vocal articulations of multiple persons, each with height h_i. We aim at inferring $\mathrm{f}(\mathbf{x})$ from $\mathbf{D}$, and anticipate that this regression model will be able to estimate the value of h_{n+1} for new unseen input $\mathbf{x}_{n+1}^{(d)}$, which is generated from the same underlying process that generated $\mathbf{D}$. To do that we define $h=\mathrm{f}(\mathbf{x})+\varepsilon$, where ε stands for the cumulative 'noise', which is due to the combined effect of anatomical peculiarities among human individuals with the same height, additive acoustic interference from the environment that contaminates the speech waveform, and 'instrumentation' errors related to speech acquisition, preprocessing and parameterization process. In the following, we will allow ε to be modelled as if it was Gaussian with variance σ_n^2 but this simplification is introduced mainly for simplifying the model estimation, and therefore we do not expect that it will hold true in the general case. Another simplification here is that we assume ε independent of $\mathrm{f}(\mathbf{x})$.

In brief, for a Gaussian process [12] obtained from a Bayesian linear regression model $\mathrm{f}(\mathbf{x})=\phi(\mathbf{x})^T\mathbf{w}$ with prior $\mathbf{w}\sim\mathrm{N}(0,\Sigma_p)$, once we know the mean value and the covariance matrix of the model, we can compute the mean and covariance for an unlabeled input $\mathbf{x}_{n+1}$, such as

$$\mathrm{E}[\mathrm{f}(\mathbf{x})] = \phi(\mathbf{x})^T\,\mathrm{E}[\mathbf{w}],\ \text{and}$$
$$\mathrm{E}[\mathrm{f}(\mathbf{x})\,\mathrm{f}(\mathbf{x}_{n+1})] = \phi(\mathbf{x})^T\,\mathrm{E}[\mathbf{w}\mathbf{w}^T]\phi(\mathbf{x}_{n+1}) = \phi(\mathbf{x})^T\Sigma_p\phi(\mathbf{x}_{n+1}).$$

Here, E[.] is the expected value, and $\mathrm{f}(\mathbf{x})$ and $\mathrm{f}(\mathbf{x}_{n+1})$ are jointly Gaussian with zero mean and a covariance given by $\phi(\mathbf{x})^T\Sigma_p\phi(\mathbf{x}_{n+1})$. The covariance for the training set is $\mathrm{cov}(\mathbf{h})=K(\mathbf{X},\mathbf{X})+\sigma_n^2\mathbf{I}$, and the joint distribution of observing the training and test data under the prior is

$$\begin{bmatrix}\mathbf{h}\\ \mathrm{f}(\mathbf{x}_{n+1})\end{bmatrix}\sim N\left(0,\begin{bmatrix}K(\mathbf{X},\mathbf{X})+\sigma_n^2\mathbf{I} & k(\mathbf{X},\mathbf{x}_{n+1})\\ k(\mathbf{x}_{n+1},\mathbf{X}) & k(\mathbf{x}_{n+1},\mathbf{x}_{n+1})\end{bmatrix}\right).$$

Next, the predictive distribution for Gaussian process regression is $\mathrm{f}(\mathbf{x}_{n+1})\mid\mathbf{X},\mathbf{h},\mathbf{x}_{n+1}\sim N\left(\bar{\mathrm{f}}(\mathbf{x}_{n+1}),\mathrm{cov}(\mathrm{f}(\mathbf{x}_{n+1}))\right)$, where $\bar{\mathrm{f}}(\mathbf{x}_{n+1})=\mathrm{k}(\mathbf{x}_{n+1})^T\left(\mathrm{K}(\mathbf{X},\mathbf{X})+\sigma_n^2\mathbf{I}\right)^{-1}\mathbf{h}$, and $\mathrm{cov}(\mathrm{f}(\mathbf{x}_{n+1}))=k(\mathbf{x}_{n+1},\mathbf{x}_{n+1})-k(\mathbf{x}_{n+1})^T\left(K(\mathbf{X})+\sigma_n^2\mathbf{I}\right)^{-1}k(\mathbf{x}_{n+1})$. In this manner, for a spoken utterance which is parameterized by the vector $\mathbf{x}_{n+1}^{(d)}$, the model, $\mathrm{f}(\mathbf{x})$, estimates not only the height h_{n+1} for unseen speakers but also the variance of this estimation, which indicates the degree of uncertainty.

The model creation steps are summarized in Figure 1, and the height estimation process is illustrated in Figure 2. As Figure 1 presents, during training speech utterances are parameterized to a set of speech features. The speech features are afterwards ranked with respect to their relevance to the height estimation problem, and a subset of them are selected for the feature vector. Finally, a regression model is created using the feature vectors and their corresponding ground truth labels, h_i. The regression model and the list of indexes of the relevant speech features are stored for further use. During the operation of the height estimator (refer to Figure 2), a multi-sensor microphone array acquires the audio waveforms. The multi-channel audio is amplified, sampled, quantized by level, and next is converted to a single-channel audio signal. (Alternatively, single sensor audio acquisition is also applicable if speaker localization and tracking are not needed). After noise reduction and sound source enumeration the single speaker utterances are kept. A speech activity detector is used for eliminating the silence portions of the waveform, and only the speech segments are parameterized. The final feature vector is formed by retrieving only those speech features, which were selected as beneficial during the training stage. Finally, the regression model, $\mathrm{f}(\mathbf{x})$, estimates the height, h_{n+1}, that corresponds to the present input feature vector, $\mathbf{x}_{n+1}^{(d)}$, as well as the uncertainty of this estimation.

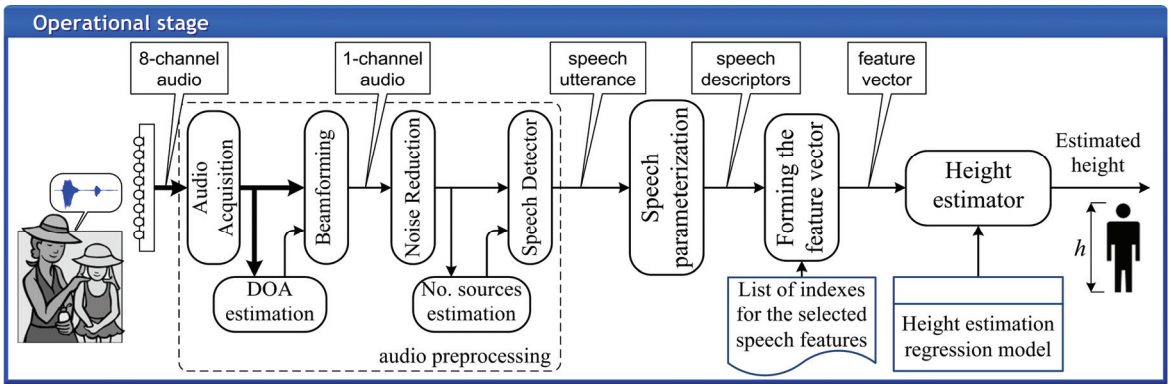


Figure 2 – Regression-based scheme for height estimation from speech for unknown speakers: the operational mode.

3. IMPLEMENTATION

For the purpose of model development, we made use of the TIMIT database [13]. The training set consists of 462 speakers, including 326 males and 136 females, and the test subset consists of the recordings of 168 speakers, including 112 males and 56 females. In both subsets, each speaker utters 10 utterances. Likewise previous related work [4, 7], in the following experiments the speaker MCTW0 was excluded from the test subset, since his height of 2.032 meters is out of the range of heights represented in the training subset, $h_{trn} \in [1.448, 1.981]$. The test set was used for the purpose of model validation. We have downsampled both training and test subsets to 8 kHz, with 16 bits per sample.

3.1 Feature Ranking and Selection

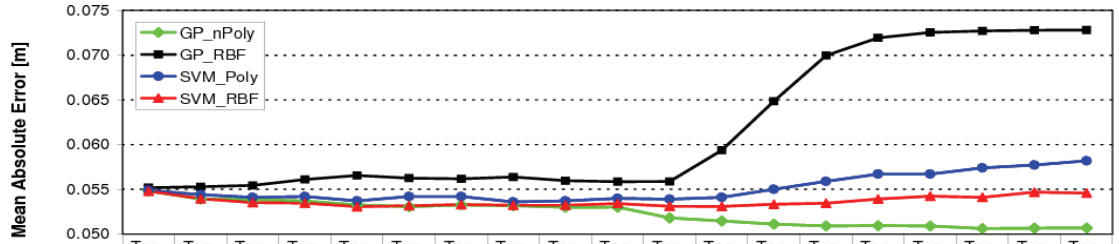
Numerous studies on the relevance of the basic speech parameters (speech energy, pitch, formants, MFCC, LPC, etc), with respect to the height estimation problem, led to the conclusion that specific speech descriptors are weakly or moderately correlated with the human height and that some are better correlated with height than others. Since, there is no single speech feature or a small set of speech features, which would permit an accurate estimation of the human height, we hypothesize that an automatic height estimator would rely on a large set of speech descriptors, which hopefully will be complementary to each other, and when combined would contribute to increase of the overall estimation accuracy.

In the present work, we made use of the openSMILE [9] audio parameterization framework, which computes 6552 utterance-level audio descriptors. These are statistical parameters, which are computed on the basic frame-level audio descriptors, such as the root mean square (RMS) frame energy, the zero-crossing rate (ZCR) from the time-domain speech signal, the harmonics-to-noise ratio (HNR) by autocorrelation function, the pseudo loudness, the Mel-spectra, the twelve MFCCs, the fundamental frequency of speech normalized to 500 Hz, the voice quality etc, and their first and second time-derivatives. Among the statistical functional parameters that form the utterance-level feature vector are the mean, standard deviation, kurtosis, skewness and higher order moments, segments, extreme values (minimum, maximum, relative position and range), linear and quadratic regression coefficients (offset, slope, mean square error, etc), percentiles, durations, onsets, DCT coefficients etc, which are computed over the basic frame-level audio descriptors.

However, the use of multidimensional feature vectors that consist of a large number of audio parameters is costly in terms of training and operational complexity, memory demands and required training data. Therefore, it is desired to reduce the dimensionality of the feature vectors by discarding audio features that are not relevant to the speaker's height and are redundant to others. To investigate the relevance of the openSMILE audio parameters, we performed ranking with the Regression Relief-F feature-ranking algorithm [14], which estimates the quality of each audio feature according to its ability to distinguish between instances, which are close to each other. In total, 3029 out of the 6552 audio features were found to be to some degree relevant to the height estimation problem, i.e. they demonstrated positive attribute quality value. Next, the ordered list of attributes, which resulted from the feature ranking, served for the selection of subsets of various sizes, which consist of the top-n ranked audio descriptors, with $n \in \{1, 2, \ldots, 10, 20, \ldots, 100, 200, \ldots, 1000\}$. The appropriateness of these subsets was evaluated through measuring the accuracy of height estimation for two GP-based and two support vector machine (SVM)-based regression models. Here the SVM models are considered the baseline, as they are known to offer state-of-the-art performance and to cope well with high dimensional feature space.

As presented in Figure 3, the GP-based model implemented with normalized polynomial kernel, GP_nPoly, presented the best overall accuracy and is competitive to the SVM based regression model. When the size of the feature vector increases over the top-200 attributes, some increase of the mean absolute error (MAE) is observed for the SVM with RBF kernel and with polynomial kernel and much larger increase for the GP with RBF kernel. However, the accuracy for the GP-based model with normalized polynomial kernel, GP_nPoly, improved slightly, mainly due to the advantages of the normalized polynomial kernel. In the following we detail only on the results for GP_nPoly.

Although the best accuracy for the GP_nPoly model is observed for large feature vectors (top-300 to top-1000), due to practical reasons in the following we will rely on a feature vector composed of the top-50 attributes, as this set offers a reasonable trade-off between computational complexity and height estimation accuracy. Among the top-50 ranked attrib-



	Top-10	Top-20	Top-30	Top-40	Top-50	Top-60	Top-70	Top-80	Top-90	Top-100	Top-200	Top-300	Top-400	Top-500	Top-600	Top-700	Top-800	Top-900	Top-1000
GP_nPoly	0.055	0.054	0.054	0.054	0.053	0.053	0.053	0.053	0.053	0.053	0.052	0.051	0.051	0.051	0.051	0.051	0.051	0.051	0.051
GP_RBF	0.055	0.055	0.055	0.056	0.057	0.056	0.056	0.056	0.056	0.056	0.056	0.059	0.065	0.070	0.072	0.073	0.073	0.073	0.073
SVM_Poly	0.055	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.055	0.056	0.057	0.057	0.057	0.058	0.058
SVM_RBF	0.055	0.054	0.054	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.054	0.054	0.054	0.055	0.055

Figure 3 – The height recognition accuracy (MAE in meters) for various sizes of the feature vector: TIMIT test dataset.

utes multiple audio features related to the fundamental frequency (F_0) and the MFCCs were found relevant to the height of the speaker. The mean of the fundamental frequency was not ranked in the top-50 audio features, which is in agreement with previous research. However, a number of derivatives of the fundamental frequency of speech were found important for estimating the height of a person. In the top-10 ranked attributes there are three F_0-related parameters, and in the top-50 there are sixteen [15]. The last makes the F_0-based statistical parameters quite important, when compared to other basic features. The good relevance of the MFCC, reported in earlier work [4] was also confirmed because 25 out of the top-50 attributes are based on the MFCC.

3.2 Height Estimation Accuracy in Controlled Conditions

The TIMIT database [13] had been recorded in a controlled setup: American English, read sentences, constrained vocabulary, high SNR, small variations in the speaker-microphone distance etc, which guarantees controlled mismatch between training and test data. As Figure 3 shows, in these controlled conditions, we observed MAE of 0.053 meters for the top-50 feature vector. This corresponds to an averaged relative error of 3.0% with respect to the average height of 1.75 meters of the speakers in the TIMIT test data.

4. EVALUATION IN REAL-WORLD SETUP

4.1 The PROMETHEUS Database

The multimodal PROMETHEUS database [16] was created in support of RTD activities aiming at the creation of a framework for monitoring and interpretation of human behaviours in unrestricted indoor and outdoor environments. The audio part of this database consists of four hours of recordings, representative for two application scenarios: smart-home (indoors, Greek language) and public security – airport and ATM (outdoors, English spoken by non-native speakers). Each recording session is comprised of multiple action scenes concatenated in a single sequence, where each action scene is implemented a number of times by different actors. The indoor scenes were implemented by five skilled actors: three females and two males, with heights in the range [1.60, 1.72] meters. In addition, twelve supernumerary actors (including one female) were involved in the outdoor scenes. The heights of the people involved in the outdoor episodes were in the range [1.60, 1.85] meters. The actors' age was in the range [22, 56] years with mean value of 33.8 years. In the following, we report results on the so-called *selected scenes,* which the PROMETHEUS consortium identified as the most interesting from application point of view. These thirty-two scenes with individual durations between 15 and 134 seconds have a cumulative length of approximately 30 minutes and represent typical multiple-person interaction episodes.

The audio, recorded with an eight-channel uniform linear array with spacing between the microphones of 0.1 meters, is sampled at 32 kHz with resolution 32-bits per sample.

4.2 Experimental Setup

In all experiments we used off-line processing of the PROMETHEUS selected scenes described in Section 4.1. Since here we are interested only in the speech portions of the signal, we converted the 8-channel audio to sampling frequency of 8 kHz and resolution 16-bit. A minimum variance beamformer with processing window of 0.064 seconds and overlap 0.032 seconds was used. Noise reduction was performed with a band-pass Butterworth filter of order six, with low and high cut frequencies $f_{lo} = 250$ Hz and $f_{hi} = 3700$ Hz, respectively.

Instead of relying on the sound source enumeration and speech/non-speech detection components, in the present evaluation we made use of the manual segmentation of the PROMETHEUS audio data, which provides error-free segmentation of single-speaker utterances. The last evades the dependence of the height estimation accuracy on the accuracy of these two components. In total, eighty single-speaker utterances, with SNR in the range between 10 and 20 dB, were available in the twelve indoor selected scenes, and ninety-two, with SNR between 6 and 12 dB, in the twenty outdoor selected scenes. Two additional outdoor sets, designated as *lowSNR* and *wInterf*, were formed by thirty-one single-speaker segments with SNR between 0 and 3 dB and twelve speech segments with one or more concurrent speakers. The test datasets obtained to this end consisted of utterances from three females and two males for the indoors subset, and by three females and seven males for the outdoors subsets. Next, the energy detector of openSMILE was used to discard silences, and the openSMILE parameters were computed only for the speech segments. The top-50 audio features were kept for the feature vector.

The WEKA [17] implementation of GP-based regression with normalized polynomial kernel was used. The GP-based

model created from the TIMIT training set, used in Section 3, was reused in all experiments on the PROMETHEUS data.

4.3 Experimental Results

The experimental results for the GP-based model with normalized polynomial kernel are shown in Tables 1 and 2. Specifically, the height estimation accuracy for the twelve indoor and twenty outdoor scenes is reported in Table 1 in terms of MAE, root mean square error (RMSE), and averaged relative error (ARE) with respect to the average height μ_t, and in Table 2 in terms of percentages of height estimations with error within a specific range. In Table 1, we also show the average of ground truth heights, μ_t, the average of the estimated heights, μ_e, the difference μ_t-μ_e which indicates the bias of the estimations for the specific dataset, and the standard deviation of the height estimation error, σ_e.

As the tables show, despite the mismatch between the training (TIMIT) and test (PROMETHEUS selected scenes) conditions, the proposed height estimator performed well both in the indoor and in the outdoor scenes, especially in the cases of reasonable SNR (i.e. SNR between 6 dB and 20 dB). The observed MAE of 0.041 and 0.050 meters for the indoor and outdoor scenes, are comparable to the 0.053 meters for the test dataset of TIMIT, and show that to some degree the proposed height estimation scheme is robust against environmental noise. The above mentioned MAEs correspond to AREs of 2.5%, 3.0% and 3.0%, respectively (Table 1).

Next, Table 2 shows that the increase of MAE and ARE for the *lowSNR* dataset is mostly due to the increased value of the estimation error, and not much to the increased quantity of errors. However, in the case of concurrent speakers, as in dataset *wInterf*, the error exceeded 0.05 meters for nearly all estimations, i.e. concurrent speech was found devastating to the height estimation accuracy. (However, we also acknowledge that portion of the error could be due to the manner in which ground truth heights were set in the segments with concurrent speakers: We selected the height of the speaker with the longer speech activity as the ground truth.)

As Table 1 shows, there is some correlation between the sign of the error (μ_t–μ_e) and the condition indoors/outdoors. Specifically, in the indoor condition (including TIMIT) the height estimator tends to overrate the speaker's height, while outdoors the speakers are underrated. We deem that this phenomenon might be linked to the difference in the reverberation times and the acoustics between indoor/outdoor conditions. However, we admit that a further in-depth study is required for better understanding of this phenomenon.

Finally, although we find the proposed height estimation scheme appropriate for use in unconstrained conditions, and deem it is applicable to real-world surveillance applications, we admit that further studies on the sensitivity of the height estimation with respect to the accuracy of the single-speaker utterance detection and the speech/non-speech separation (cf. Figure 2) are required.

Table 1. Height estimation accuracy in terms of MSE, RMSE and ARE. All values are in meters, except for the AREs.

	μ_t	μ_e	μ_t-μ_e	σ_e	*MAE*	*RMSE*	*ARE*
Indoor, [10, 20]dB	1.679	1.695	-0.016	0.054	0.041	0.056	2.5%
Outdoor, [6, 12]dB	1.678	1.648	0.030	0.054	0.050	0.062	3.0%
Outd., *lowSNR*, [0, 3]dB	1.743	1.723	0.020	0.090	0.067	0.093	3.9%
Outd., *wInterf*, 0dB	1.707	1.640	0.067	0.083	0.097	0.107	5.7%
TIMIT test dataset	1.750	1.750	-1.7e-4	0.068	0.053	0.067	3.0%

Table 2. Height estimation accuracy – each cell shows the percentage of height estimations within the specific error range

Error range [m] ≤	0.01	0.02	0.025	0.05	0.075	0.10	0.125	0.15
Indoor, [10, 20]dB	15.0	33.8	43.8	72.5	82.5	95.0	97.5	97.5
Outdoor, [6, 12]dB	16.3	22.8	27.2	54.3	79.3	91.3	96.7	98.9
Outd., *lowSNR*, [0, 3]dB	16.1	29.0	35.5	51.6	64.5	71.0	80.6	87.1
Outd., *wInterf*, 0dB	0.0	0.0	0.0	8.3	50.0	66.7	75.0	83.3

ACKNOWLEDGEMENTS

This research was supported by the PROMETHEUS project (www.prometheus-fp7.eu), which is co-funded by the FP7 of the European Union under Grant Agreement FP7-ICT-214901.

REFERENCES

[1] A. K. Jain, S. C. Dass, and K. Nandakumar, "Can soft biometric traits assist user recognition?" In *Biometric Technology for Human Identification*. Eds. A.K. Jain and N. Ratha, vol. 5404, pp. 561–572.

[2] I. Kispál and E. Jeges, "Human height estimation using a calibrated camera," In *Proc. CVPR 2008*.

[3] A. Gallagher, A. Blose, T. Chen, "Jointly Estimating Demographics and Height with a Calibrated Camera", In *Proc. ICCV09*.

[4] B. L. Pellom and J. H. L. Hansen, "Voice analysis in adverse conditions: the centennial Olympic park bombing 911 call," In *Proc. MWSCAS 1997*, vol. 2, pp. 873–876.

[5] L. H. Smith and D. J. Nelson, "An estimate of physical scale from speech," In *Proc. ICASSP 2004*, vol. 1, pp. 561–564.

[6] M. Blomberg and D. Elenius, "Estimating speaker characteristics for speech recognition," In *Proc. FONETIK 2009*, pp. 154–158.

[7] S. Dusan, "Estimation of speaker's height and vocal tract length from speech signal," In *Proc. Interspeech 2005*, pp. 1989–1992.

[8] I. Mporas and T. Ganchev, "Estimation of unknown speaker's height from speech", *International Journal of Speech Technology*, vol. 12, no. 4, 2009.

[9] F. Eyben, M. Wöllmer, and B. Schüller, "openEAR – introducing the Munich open-source emotion and affect recognition toolkit," In *Proc. ACII-2009*, IEEE, Amsterdam, The Netherlands, 2009.

[10] W. T. Fitch and J. Giedd, "Morphology and development of human vocal tract: a study using magnetic resonance imaging," *JASA*, vol. 106, no. 3, pp. 1511–1522, 1999.

[11] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[12] D. J. C. MacKay, *Introduction to Gaussian Processes*. Department of Physics, Cambridge University, UK. 1998.

[13] J. Garofolo, "Getting started with the DARPA-TIMIT CD-ROM: an acoustic phonetic continuous speech database", NIST, Gaithersburgh, MD, USA, 1988.

[14] M. Robnik-Šikonja and I. Kononenko, "An adaptation of Relief for attribute estimation in regression," In *Proc. ICML-1997*, pp. 296–304.

[15] T. Ganchev, I. Mporas, and N. Fakotakis, "Audio Features Selection for Automatic Height Estimation from Speech", In *Proc. SETN 2010*, LNAI 6040/2010. Springer-Verlag, pp. 81–90.

[16] S. Ntalampiras, D. Arsić, A. Störmer, T. Ganchev, I. Potamitis, and N. Fakotakis, "Prometheus database: a multimodal corpus for research on modeling and interpreting human behavior", In *Proc. DSP-2009*, Santorini, Greece, 2009.

[17] H. I. Witten and E. Frank, *Data Mining: practical machine learning tools and techniques*, Morgan Kaufmann Publishing, 2005.