Height Estimation from Speech Signals using i-vectors and Least-Squares Support Vector Regression

Amir Hossein Poorjam, Mohamad Hasan Bahari, Vasileios Vasilakakis, and Hugo Van hamme

***Abstract*—This paper proposes a novel approach for automatic speaker height estimation based on the i-vector framework. In this method, each utterance is modeled by its corresponding i-vector. Then artificial neural networks (ANNs) and least-squares support vector regression (LSSVR) are employed to estimate the height of a speaker from a given utterance. The proposed method is trained and tested on the telephone speech signals of National Institute of Standards and Technology (NIST)2008 and 2010 Speaker Recognition Evaluation (SRE) corpora respectively. Evaluation results show the effectiveness of the proposed method in speaker height estimation.**

***Keywords*—Artificial Neural Networks, i-vector, Least-squares Support Vector Regression, Speaker Height Estimation.**

I. Introduction

In many forensic cases, evidence might be in the form of voice recordings, e.g. a threat call and a blackmail call. Forensic experts might have a list of suspects but it can take time to check them all. In such cases, it could be beneficial to rank them according to objective criteria such as gender, age and accent in order to narrow down the number of suspects [1]–[3]. In this paper, we focus on speaker height estimation.

Experimental studies have found different acoustic cues for speaker height estimation [4], [5]. However, the relation of these acoustic cues with speaker age is usually complex and affected by many other factors such as speech content, language, gender, weight, emotional condition, smoking and drinking habits. Furthermore, in many practical cases we have no control over the available speech duration, content, language, environment, recording device and channel conditions. Therefore, height estimation from speech signals is a very challenging task.

Previous studies have investigated a correlation between the speech signal of a person and his/her height. In experiments conducted by Van Dommelen and Moxness, the ability of listeners to estimate the height of speakers from their voice have been examined. In this study, significant correlations between estimated and actual height of male speakers were reported [4]. In studies on speech-driven automatic height estimation, several resources have been devoted to identify acoustic features of speech that can convey information about speaker height. For example, [4] and [5] analyzed the correlation between speaker height and formant frequencies, based on the assumption of speech production theory that there is a correlation between a person's vocal tract length (VTL) and his/her height. Recently, Arsikere et al. proposed a new algorithm based on the assumption of the uniform tube model of the subglottal system to estimate the speakers' height from the subglottal resonances (SGRs) [6], [7]. In other studies, Pellom and Hansen performed height group recognition by applying Mel-frequency cepstral coefficients (MFCCs) to train a height-dependent Gaussian mixture model. Then a maximum a posteriori classification rule was used to assign each audio file to one of several height groups [8]. However, this text independent approach does not estimate the actual height of a speaker, which can be achieved by using regression techniques. Ganchev et al. applied a large set of openSmile audio descriptors and performed support vector regression to estimate the height of a test speaker [9].

In this paper we suggest a speech-based automatic height estimation method. We propose a new method for automatic height estimation based on i-vectors instead of raw acoustic features as in previous studies. In the field of speaker recognition, recent advances using the i-vector framework [10] have increased the classification accuracy considerably. An i-vector is a compact representation of an utterance in the form of a low-dimensional feature vector.

To select an accurate regression approach for this problem, two different function approximation approaches, namely least squares support vector regression (LSSVR) and artificial neural networks (ANNs) are compared. We also investigate the effect of the kernel in LSSVR and of the training algorithm in ANN. Evaluation on the NIST 2008 and 2010 SRE corpora shows the effectiveness of the proposed approach.

The rest of the paper is organized as follows. In the section II the problem of automatic height estimation is formulated and the proposed approach is described. Section III explains our experimental setup. The evaluation results are presented and discussed in Section IV. The paper ends with conclusions in section V.

Manuscript received February 14, 2014.

A.H. Poorjam is with the Center for Processing Speech and Images, KU Leuven, Belgium, (corresponding author, phone: +32-16-328545; fax: +32-16-321723; e-mail: amir.poorjam@student.kuleuven.be).

M.H. Bahari is with the Center for Processing Speech and Images, KU Leuven, Belgium (e-mail: MohamadHasan.Bahari@esat.kuleuven.be).

V. Vasilakakis is with Polytechnic University of Turin, Italy (e-mail: Vasileios.vasilakakis@polito.it).

H. Van hamme is with the Center for Processing Speech and Images, KU Leuven, Belgium (e-mail: Hugo.vanhamme@esat.kuleuven.be).

II. SYSTEM DESCRIPTION

A. Problem Formulation

In the speaker height estimation problem, we are given a set of training data $D = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^p$ denotes the i^{th} utterance and $y_i \in \mathbb{R}$ denotes the corresponding height. The goal is to design an estimator function g, such that for an utterance of an unseen speaker $\mathbf{x}_{tst}$, the estimated height $y = g(\mathbf{x}_{tst})$ approximates the actual height as good as possible in some predefined sense.

B. Height Estimation Using i-vectors

The first step for approximating function g is converting variable-duration speech signals into fixed-dimensional vectors suitable for regression algorithms. In this research, we apply the i-vector framework for this purpose. i-vector based techniques have recently been effectively applied to speaker verification and recognition [10], language recognition [11], speaker age estimation [12] and accent recognition [13]. The i-vector framework, which is also referred to as total variability modeling, assumes the GMM mean supervector $\boldsymbol{\mu}$ can be decomposed as

$$\boldsymbol{\mu} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \tag{1}$$

where $\boldsymbol{m}$ is the universal background model (UBM) mean supervector, with C mean components of dimension F. Subspace matrix T denotes a skinny matrix of size $C.F \times M$. $\boldsymbol{w}$ is a latent vector of size M, which is referred to as i-vector. An efficient Maximum-Likelihood estimate of matrix $\boldsymbol{T}$ and a Maximum-a-posteriori (MAP) estimation of $\boldsymbol{w}$ considering prior standard normal distribution $N(0, I)$ can be found in [14].

C. Function Approximation

1) Least Squares Support Vector Regression: Support vector regression (SVR) is a function approximation approach developed as a regression version of the widely known Support Vector Machines (SVM) classifier [15]. Using nonlinear transformations, SVMs map the input data into a higher dimensional space in which a linear solution can be calculated. They also keep a subset of the samples which are the most relevant data for the solution and discard the rest. This makes the solution as sparse as possible. While SVMs perform the classification task by determining the maximum margin separation hyperplane between two classes, SVRs carry out the regression task by finding the optimal regression hyperplane in which most of training samples lie within an ε-margin around this hyperplane [15], [16].

In this paper, we use the least squares version of support vector regression (LSSVR). While a SVR solves a quadratic programming, which results in high algorithmic complexity and memory requirement, a LSSVR involves solving a set of linear equations [16] which speeds up the calculations. This simplicity is achieved at the expense of loss of sparseness, therefore all samples contribute to the model, and consequently, the model often becomes unnecessarily large. In this paper, linear and radial basis function (RBF) kernels are used to approximate $g(\mathbf{x})$. For the LSSVR with RBF kernels, a 5-fold cross-validation to tune the smoothing parameter of the kernels is used.

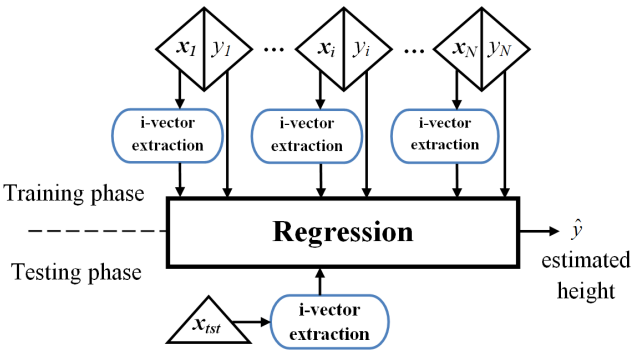

Fig. 1. Block diagram of the proposed speaker height estimation approach in training and testing phases.

2) Neural Network Regression: A multilayer perceptron (MLP) is a supervised, feed-forward neural network, which is widely applied to regression problems due to their ability to approximate complex nonlinear functions from input data [17]–[20]. An MLP usually utilizes a derivative based optimization algorithm such as back-propagation to train the network. Different training methods have been suggested during the last decades [17]–[20] to enhance the training speed, provide more memory efficient methods and represent better convergence properties. In this research, to reach an accurate network, we apply four training algorithms.

The first one, namely the Levenberg-Marquardt (LM) algorithm, uses step size damping by regularizing the Hessian matrix and exhibits a fast training [17]. In the second training approach, the search direction is computed from the new gradient and the previous search direction, based on the Fletcher-Reeves variation of the conjugate gradient method (CGF) [18]. The third technique, labeled as BFG in this paper, is a quasi-Newton method for back-propagation, that converges in few iterations but that requires more computation in each iteration [19]. The fourth training scheme is the Levenberg-Marquardt algorithm with a Bayesian regularization that minimizes a linear combination of the squared error and squared weights, such that the network will have good generalization capability [20].

D. Training and Testing

The proposed height estimation approach is depicted in Fig. 1. During the training phase, each utterance is mapped onto a 400 dimensional vector using the i-vector framework. The obtained i-vectors of the training set are then used as features with their corresponding height labels to train a regressor for approximating function g. During the testing phase, an i-vector is extracted from the test utterance and the estimated height is obtained using the trained regression function.

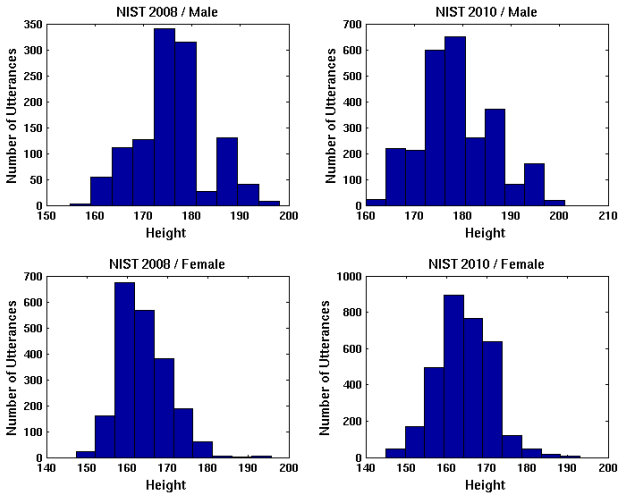

Fig. 2. The height histogram of telephone speech utterances for the NIST 2008 and NIST 2010 databases.

III. Experimental Setup

A. Database

The National Institute for Standard and Technology (NIST) have held annual or biannual speaker recognition evaluations (SRE) for the past two decades. With each SRE, a large corpus of telephone (and more recently microphone) conversations are released along with an evaluation protocol. These conversations typically last 5 minutes and originate from a large number of participants for whom additional meta data is recorded including age, height, language and smoking habits. The NIST databases were chosen for this work due to the large number of speakers and because the total variability subspace requires a considerable amount of development data for training. The development data set used to train the total variability subspace and UBM includes over 30,000 speech recordings and was sourced from the NIST 2004-2006 SRE databases, LDC releases of Switchboard 2 phase III and Switchboard Cellular (parts 1 and 2).

For the purpose of height estimation, telephone recordings from the common protocols of the recent NIST 2008 and 2010 SRE databases are used for training and testing, respectively. The core protocol, short2-short3, from the 2008 database contains 3999 telephone recordings of 1236 speakers whose height is known. Similarly, the extended core-core protocol of the 2010 database contains 5792 telephone segments from 445 speakers. The height histogram of male and female speakers of NIST 2008 and 2010 SRE databases of target are depicted in Fig. 2.

B. Performance Metric

In order to evaluate the effectiveness of the proposed system, we used the mean absolute error (MAE) of the speakers' estimated height, and the Pearson correlation coefficient (CC) between the actual speakers' height and estimated speakers' height. MAE is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|f_i - y_i| \tag{2}$$

where f_i is the i^{th} estimated height and y_i is the i^{th} actual height, and N is the total number of test samples.

Although MAE is a helpful performance metric in regression problems, it is limited in some respects specially in the case of a test set with a skewed distribution. Therefore, we use correlation coefficient, which is computed as:

$$CC = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{f_i - \bar{f}}{s_f}\right)\left(\frac{y_i - \bar{y}}{s_y}\right), \tag{3}$$

where $\bar{f}$ and s_f denote sample mean and standard deviation, respectively.

IV. Results and Discussion

In this section, the proposed speaker height estimation approach is evaluated. The acoustic feature consists of 20 Mel-Frequency Cepstrum Coefficients (MFCCs) including energy appended with their first and second order derivatives, forming a 60 dimensional acoustic feature vector. This type of feature is very common in state-of-the-art i-vector based speaker recognition systems. To have more reliable features, Wiener filtering, speech activity detection [21] and feature warping [22] have been considered in the front-end processing.

The results of using MLP trained using four different algorithms, namely LMB, CGF, BFG and BR are listed in Table I. For each training algorithm, the network was trained using different number of hidden layers, hidden neurons and activation functions. Then, based on the obtained results on the development set, the best network architecture has been selected to be evaluated on the test data. The development set consists of 25% of data of the NIST 2008 SRE database so that none of them were used in the training set. In addition, since there are several utterances from each speaker in the data set, the development set was selected such that there was no speaker who had utterances in both training and development sets.

For the three-layer NN, 10 hidden neurons and for the four-layer NN, 20 neurons in the first hidden layer and 5 neurons in the second hidden layer have been selected, respectively. The activation function for hidden layers is a logistic sigmoid function. In order to perform regression, a linear activation function has been utilized for the output layers. To attenuate the effect of random initialization, the training and testing phases of each experiment was repeated 20 times.

The evaluation results mentioned in Table I have been reported by averaging the performances over all 20 experiments. As reported in Table I, a MLP with the BFG training algorithm yields more accurate height estimation results compared the rest of training methods for both male and female speakers.

The results of using LSSVR as a function approximation method are listed in Table II. In this paper two different kernels, namely linear kernel and radial basis function (RBF) kernel have been used to approximate the function g. The

TABLE I
SPEAKER HEIGHT ESTIMATION USING A MLP WITH DIFFERENT TRAINING ALGORITHMS. CC IS THE PEARSON CORRELATION COEFFICIENT BETWEEN ACTUAL AND ESTIMATED HEIGHT.

Training Algorithm	Male		Female	
	Three Layers	Four Layers	Three Layers	Four Layers
LMB	0.23	0.25	0.20	0.20
CGF	0.35	0.36	0.33	0.32
BFG	**0.35**	**0.36**	**0.36**	**0.35**
BR	0.34	0.24	0.27	0.23

hyper-parameters of the RBF kernel have been tuned using a 5-fold cross-validation. After optimization of the hyper parameters, the model has been trained.

As it is shown in Table II, the linear kernel is more effective than the RBF kernel in this problem. In this case, CC for male speakers, female speakers and when the male and female data were pooled together are 0.40, 0.41 and 0.60 respectively. The scatter plots of estimation for male speakers, female speakers and when the male and female data were pooled together are shown in Fig. 3 and 4 respectively. The mean absolute error (MAE) of estimation is 6.2 cm and 5.8 cm of for male and female speakers respectively. Although the obtained MAE is satisfactory and the correlation coefficient is fairly strong when male and female data are pooled together, the CC within male and female speakers requires improvement. Unfortunately there is no published results on the same database for comparison purpose. However, the results of published papers on other datasets indicate the typical range of performance in automatic speaker height estimation problem. In [7], reported CC of speaker height estimation on TIMIT database using a method based on sub-glottal resonances [6] are 0.12, 0.21 and 0.71 for male speakers, female speakers and when the male and female data were pooled together respectively. In [8], the obtained CC of speaker height estimation for male and female speakers of TIMIT database using a GMM based approach are 0.39 and 0.31 respectively. The obtained results seem to be reasonable, considering that the applied testing dataset in this paper consists of spontaneous telephone speech signals and the number of test set speakers in this paper (3999 telephone recordings of 1236 speakers) is considerably larger than that of [7] and [8].

V. CONCLUSIONS

In this paper, utterance modeling with i-vectors has been used in conjunction with an ANN and LSSVR to address speaker height estimation. To evaluate the proposed estimator, telephone utterances of the NIST 2008 and 2010 SRE databases were used for training and testing respectively.

TABLE II
SPEAKER HEIGHT ESTIMATION USING LSSVR WITH DIFFERENT KERNELS. CC IS THE PEARSON CORRELATION COEFFICIENT BETWEEN ACTUAL AND ESTIMATED HEIGHT.

Kernel type	Male	Female
Linear	**0.41**	**0.40**
RBF	0.30	0.23

Experimental results show the effectiveness of the proposed approach. The obtained results also show that a LSSVR with a linear kernel is more accurate than several architecture of ANN and a LSSVR with a RBF kernel in this problem.

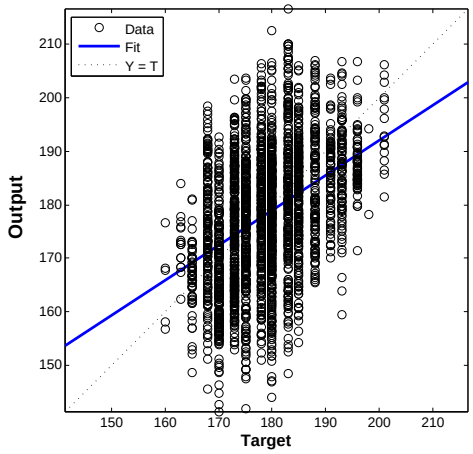

Fig. 3. The scatter plot of height estimation for male speakers.

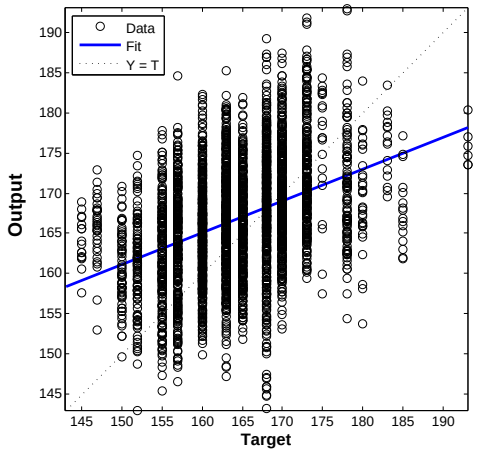

Fig. 4. The scatter plot of height estimation for female speakers.

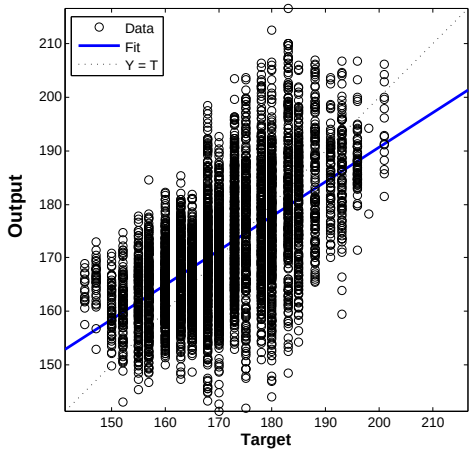

Fig. 5. The scatter plot of height estimation for both male and female speakers.

REFERENCES

[1] M. H. Bahari and H. Van hamme, "Speaker age estimation and gender detection based on supervised non-negative matrix factorization," in *Proc. IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications*, 2011, pp. 1–6.

[2] M. H. Bahari *et al.*, "Speaker age estimation using hidden markov model weight supervectors," in *11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, 2012, pp. 517–521.

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

[4] W. A. Van Dommelen and B. H. Moxness, "Acoustic parameters in speaker height and weight identification: sex-specific behaviour," *Language and Speech*, vol. 38, pp. 267–287, 1995.

[5] J. Gonzalez, "Formant frequencies and body size of speaker: a weak relationship in adult humans," *Journal of Phonetics*, vol. 32, pp. 277–287, 2004.

[6] *Automatic height estimation using the second subglottal resonance*. Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on Acoustics, Speech and Signal Processing, 2012.

[7] H. Arsikere, G. K. Leung, S. M. Lulich, and A. Alwan, "Automatic estimation of the first three subglottal resonances from adults speech signals with application to speaker height estimation," *Speech Communication*, vol. 55, no. 1, pp. 51–70, 2013.

[8] B. L. Pellom and J. H. L. Hansen, "Voice analysis in adverse conditions: the centennial olympic park bombing 911 call," in *Proc. Of the 40th Midwest symposium on circuits and systems*, 1997.

[9] T. Ganchev, I. Mporas, and N. Fakotakis, "Audio features selection for automatic height estimation from speech," *Artificial Intelligence: Theories, Models and Applications Lecture Notes in Computer Science*, vol. 6040, pp. 81–90, 2010.

[10] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front–end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[11] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via ivectors and dimensionality reduction," in *Proc. Interspeech*, 2011.

[12] M. H. Bahari, M. McLaren, H. Van hamme, and D. Van Leeuwen, "Age estimation from telephone speech using i-vectors," in *INTERSPEECH*, 2012, pp. 506–509.

[13] M. H. Bahari, R. Saeidi, H. Van hamme, and D. Van Leeuwen, "Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7344–7348.

[14] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verication," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[15] A. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[16] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least squares support vector machines*. World Scientific, 2002.

[17] M. T. Hagan and M. Menhaj, "Training feed-forward networks with the marquardt algorithm," *EEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[18] L. E. Scales, *Introduction to Non-Linear Optimization*. Springer-Verlag, 1985.

[19] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*. Emerald, 1981.

[20] D. J. C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[21] M. McLaren and D. van Leeuwen, "A simple and effective speech activity detection algorithm for telephone and microphone speech," in *Proc. NIST SRE Workshop*, 2011.

[22] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," pp. 213–218, 2001.