A Novel Approach to Speaker Weight Estimation Using a Fusion of the i-vector and NFA Frameworks

Poorjam, Amir Hossein; Bahari, Mohamad Hasan; Van hamme, Hogo

Published in:
Journal of Electrical Systems and Signals

DOI (link to publication from Publisher):
10.22067/ess.v3i1.52091

Publication date:
2017

Document Version
Accepted author manuscript, peer reviewed version

A Novel Approach to Speaker Weight Estimation Using a Fusion of the i-vector and NFA Frameworks

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

Abstract. This paper proposes a novel approach for automatic speaker weight estimation from spontaneous telephone speech signals. In this method, each utterance is modeled using the i-vector framework which is based on the factor analysis on Gaussian Mixture Model (GMM) mean supervectors, and the Non-negative Factor Analysis (NFA) framework which is based on a constrained factor analysis on GMM weight supervectors. Then, the available information in both Gaussian means and Gaussian weights is exploited through a feature-level fusion of the i-vectors and the NFA vectors. Finally, a least-squares support vector regression is employed to estimate the weight of speakers from the given utterances.

The proposed approach is evaluated on spontaneous telephone speech signals of National Institute of Standards and Technology 2008 and 2010 Speaker Recognition Evaluation corpora. To investigate the effectiveness of the proposed approach, this method is compared to the i-vector-based speaker weight estimation and an alternative fusion scheme, namely the score-level fusion. Experimental results over 2339 utterances show that the correlation coefficients between the actual and the estimated weights of female and male speakers are 0.49 and 0.56, respectively, which indicate the effectiveness of the proposed method in speaker weight estimation.

Keywords: I-vector, least-squares support vector regression, non-negative factor analysis, speaker weight estimation.

1. Introduction

The voice of a speaker conveys information about speaker's traits and states such as age, gender, body size (weight/height) and emotional state [1], [2]. Estimation of speaker's weight (which is considered as a long term trait of a speaker and an important parameter in various applications) is an interesting and challenging task in forensic, medical and commercial applications. In forensic scenarios, body size estimation of suspects from their voices can direct investigations to find cues in judicial cases. In service customization, automatic weight estimation may help users to receive services proportional to their physical conditions.

The relation between the size of various components of the sound production system (such as vocal folds and vocal tract) and the body size of a speaker has motivated researchers in the field of speaker recognition to look for features of an acoustic signal that provide cues to the body size of speakers. For instance, authors in [3] found a relationship between formants and the length of the vocal tract, based on the source-filter theory. Thus, since the vocal tract is a part of speaker's body, this feature can be used to estimate the weight of a speaker [4].

However, speaker weight estimation from the voice patterns is challenging. For instance, mean fundamental frequency (f_0) of voice is reported as a feature which has a (negative) correlation with body size. That is, females and children have higher f_0, while in males (who are taller and heavier), this value is lower [5]. However, when the relation of the fundamental frequency (f_0) and weight was investigated within male and female speakers, no correlation was found between f_0 and the weight of adult humans [6], [7]. The lowest fundamental frequency of voice (F_0^{min}) is another feature which is determined by the mass and length of the vocal folds [6]. By investigating this feature, researchers have found no correlation between F_0^{min} and weight in adult human speakers [6], [7]. Fitch has found formant dispersion (the averaged difference between adjacent pair of formant frequencies) a reliable feature which has a correlation with both vocal tract length and body size in macaques [8]. However, a weak relation between formant parameters and weight of human adults is reported in study conducted by Gonzalez [9]. This weak correlation may be due to the fact that the vocal folds in humans at puberty grow independent of the rest of the head and body. This issue is more evident in the males than the females [10], [11]. Gonzalez studied the correlation between formant frequencies and weight in human adults [9]. He calculated the formant parameters by means of a long-term average analysis of running speech signals uttered by 91 speakers. In this experiment, the Pearson correlation coefficients between formants and weights for male and female speakers were reported to be 0.33 and 0.34, respectively [9]. In research conducted by Van Dommelen and Moxness [12], the ability to judge the weight of speakers from their speech samples was

Manuscript received December 9, 2015; revised 27 August 2016; accepted 23 October 2016.
A. H. Poorjam is with Audio Analysis Lab, AD:MT, Aalborg University, Denmark.
M. H. Bahari and H. Van hamme are with the Center for Processing Speech and Images, KU Leuven, Belgium.
The corresponding author's email address is: ahp@create.aau.dk

investigated. They reported a significant correlation between estimated weight (judged by listeners) and actual weight of only male speakers. In addition, they performed a regression analysis involving several acoustic features such as f_0, formant frequencies, energy below 1 kHz, and speech rate. The results showed that the speech rate was the only parameter which had a significant correlation with male speaker's weight. They concluded that speech rate of male speakers is a reliable predictor for weight estimation.

Modeling speech utterances with Gaussian mixture model (GMM) mean supervectors is demonstrated to be an effective approach to speaker recognition [13]. However, GMM mean supervectors are high dimensional vectors, and obtaining a reliable model is difficult when limited data are available. Recently, utterance modeling using the i-vector framework [14] has considerably increased the accuracy of the classification and regression problems in the field of speaker characterization [15], [16], speaker verification [14] and language recognition [17], [18]. The i-vector, which is based on the factor analysis on GMM mean supervectors, represents an utterance in a compact and a low-dimensional feature vector. In addition, various studies show that although GMM weights convey less information than GMM means, they provide complementary information to GMM means [19]–[23]. A Non-negative Factor Analysis (NFA) framework [21], which is based on a constrained factor analysis for GMM weights, has been recently introduced and yields a new low-dimensional utterance representation. In [24], we successfully applied a score-level fusion of the i-vector and the NFA frameworks to simultaneously estimate various characteristics of speakers from speech signals. We showed that utilizing information in both GMM means and GMM weights through a score-level fusion of the i-vector and the NFA frameworks is an effective approach to improve the estimation accuracy. However, the fusion at score level requires a large development data set to train the fusion model, which results in decreasing the number of available training data.

In this study, a new speech-based method for automatic weight estimation is proposed in which instead of using raw acoustic features, each utterance is modeled using a fusion of the i-vector and the NFA frameworks at feature level. In this new utterance modeling approach, in addition to exploiting the available information in GMM means and GMM weights, the need for assigning a considerable amount of training data for development set is eliminated and speaker weight estimation is performed in one learning phase. To perform function approximation, a least-squares support vector regression (LSSVR) is utilized in this paper. For the comparison purpose, the proposed method is compared to the i-vector-based speaker weight estimation and speaker weight estimation using an alternative fusion scheme, namely the score-level fusion. The proposed approach is evaluated on spontaneous telephone speech signals of the NIST 2008 and 2010 SRE corpora. Experimental results confirm the effectiveness of the proposed approach in automatic speaker weight estimation.

The rest of the paper is organized as follows. In Section 2 the problem of automatic weight estimation is formulated and different baseline systems for speaker weight estimation are described. The proposed approach is elaborated in Section 3. Section 4 explains the experimental setup. The evaluation results are presented and discussed in Section 5. The paper ends with conclusions in Section 6.

2. Automatic Weight Estimation from Speech Signals

In this section, the problem of automatic weight estimation is formulated and different baseline approaches are described.

A. Problem Formulation

In the speaker weight estimation problem, we are given a set of training data $D = \{O_i, y_i\}_{i=1}^{N}$, where O_i denotes the i^{th} utterance and $y_i \in$ denotes the corresponding weight.

The goal is to approximate a function g, such that for an utterance of an unseen speaker, O_{tst}, the estimated weight, $\hat{y}=g(O_{tst})$, approximates the actual weight as good as possible.

B. Baseline Approaches

For the comparison purpose, the proposed method is compared to three baseline approaches, namely the basic estimator, the i-vector-based speaker weight estimation [23] and speaker weight estimation using score-level fusion of the i-vector and the NFA frameworks [24].

1) Basic Estimation System:

The output of a basic estimator is the average weight of speakers in training data set. The basic estimation system provides us a chance level accuracy.

2) The i-vector-based System:

In the i-vector-based weight estimation system, each utterance is mapped onto a 400 dimensional vector using the i-vector framework. Then, the extracted i-vectors along with their corresponding weight labels are used to train estimator. This method is considered as the baseline system in this paper.

3) The Score-Level Fusion System:

The fusion of the i-vector and the NFA frameworks is an effective approach to exploit the available information in both GMM means and GMM weights and consequently to enhance the estimation accuracy. In this method, each utterance is converted to an i-vector and an NFA vector. Then, the obtained vectors of the train set are employed to train the i-vector-based and the NFA-based models. In the next step, the i-vectors and the NFA vectors of development set are applied to the trained models. The outputs are then concatenated to form 2-dimensional vectors and along with the corresponding weight labels are used to train the fusion model. This system, which is labeled as the score-level fusion system in this paper, is presented to investigate the effect of different fusion schemes in speaker weight estimation.

3. System Description

A. Utterance Modeling

By fitting a GMM to acoustic features extracted from each speech signal, a variable-duration speech signal is converted into a fixed-dimensional vector which is suitable for regression

algorithms. The parameters of the obtained GMM characterize the corresponding utterance. Due to limited data, we are not able to accurately fit a separate GMM for a short utterance, especially in the case of GMMs with a high number of Gaussian components. Thus, for adapting a universal background model (UBM) to characteristics of utterances in training and testing databases, parametric utterance adaptation techniques are applied. In this paper, the i-vector and the NFA frameworks are applied to adapt UBM means and weights, respectively.

1) Universal Background Model and Adaptation:

Consider a UBM with the following likelihood function of data $O=\{\mathbf{o}_1, \ldots, \mathbf{o}_t, \ldots, \mathbf{o}_T\}$.

$$p(\mathbf{o}_t \mid \gamma) = \sum_{c=1}^{C} \pi_c p(\mathbf{o}_t \mid \mu_c, \mathbf{\Sigma}_c)$$
$$\gamma = \{\pi_c, \mu_c, \mathbf{\Sigma}_c\} \quad , \; c = 1, \ldots, C \tag{1}$$

where $\mathbf{o}_t$ is the acoustic vector at time t, π_c is the mixture weight for the c^{th} mixture component, $p(\mathbf{o}_t|\mu_c, \mathbf{\Sigma}_c)$ is a Gaussian probability density function with mean μ_c and covariance matrix $\mathbf{\Sigma}_c$, and C is the total number of Gaussian components in the mixture. The parameters of the UBM –γ– are estimated on a large amount of training data.

2) The i-vector Framework:

One effective method for speaker weight estimation involves adapting UBM means to the speech characteristics of the utterance. Then, the adapted GMM means are extracted and concatenated to form Gaussian mean supervectors. However, since Gaussian components of the UBM model are adapted independent of each other, some components are not updated in the case of limited training samples [25]. This problem can be alleviated by linking the Gaussian components together using the Joint Factor Analysis (JFA) framework [26].

In the JFA framework, each utterance is represented by a supervector **M** which is a speaker- and channel-dependent vector of dimension ($C.F$), where C is the total number of the mixture components in a feature space of dimension F. In the JFA framework, it is assumed that **M** can be decomposed into two supervectors:

$$\mathbf{M} = \mathbf{s} + \mathbf{c} \tag{2}$$

where $\mathbf{s} = \mathbf{u} + \mathbf{V}\boldsymbol{q} + \mathbf{D}\boldsymbol{r}$ is a speaker-dependent supervector and $\mathbf{c} = \mathbf{U}\boldsymbol{p}$ is a channel-dependent supervector. **s** and **c** are independent and possess normal distributions. **u** is the speaker- and channel-independent supervector, **V** defines a lower dimensional speaker subspace, **U** is a lower dimensional channel subspace, and **D** defines a speaker subspace. $\boldsymbol{q}$ and $\boldsymbol{r}$ are factors in speaker subspace, and $\boldsymbol{p}$ is a channel-dependent factor in channel subspace. The vectors $\boldsymbol{p}$, $\boldsymbol{q}$ and $\boldsymbol{r}$ are random variables with standard normal distributions *N (0, I)* which are jointly estimated.

In the JFA framework, the channel factor contains some information about speakers, which can be utilized in speaker identification. This fact resulted in proposing a new utterance modeling approach, referred to as the i-vector framework or the total variability modeling [27]. This method comprises both speaker variability and channel variability. Channel compensation procedures such as within-class covariance normalization can be further applied to compensate the residual channel effects in the speaker factor space [28].

The i-vector framework assumes that each utterance possesses a speaker- and channel-dependent GMM supervector which its mean, **M**, can be decomposed as

$$\mathbf{M} = \mathbf{u} + \mathbf{T}\mathbf{v} \tag{3}$$

where **u** is the mean supervector of the UBM, and **T** spans a low-dimensional subspace (400 dimensions in this work). In the i-vector framework, **T** and **v** are estimated using the Expectation-Maximization (EM) algorithm. In the E-step, **T** is supposed to be known, and **v** is updated. In the M-step, **v** is assumed to be known, and **T** is updated. The subspace vector **v** is treated as a hidden variable with the standard normal prior and the i-vector is its maximum-a-posteriori (MAP) point estimate which is calculated by maximization of the following auxiliary function over **v**.

$$\Psi(\gamma, \mathbf{v}) = \sum_{t=1}^{T} \sum_{c=1}^{C} \theta_{c,t} \log \pi_c p(\mathbf{o}_t \mid [\mu_c + \mathbf{T}_c \mathbf{v}], \mathbf{\Sigma}_c) N(\mathbf{v}) \tag{4}$$

where $N(\mathbf{v})$ denotes the standard normal distribution of **v**, $\mathbf{T}_c$ are the rows of the subspace matrix **T**, which correspond to the c^{th} Gaussian mean, and $\theta_{c,t}$ is the occupation count for the c^{th} mixture component and t^{th} frame. The occupation count is calculated as follows:

$$\theta_{c,t} = \frac{\pi_c p(\mathbf{o}_t \mid \mu_c, \mathbf{\Sigma}_c)}{\sum_{c=1}^{C} \pi_c p(\mathbf{o}_t \mid \mu_c, \mathbf{\Sigma}_c)} \tag{5}$$

In the E-step, the posterior distribution of **v** is Gaussian with the following mean $\mathbf{v}_\mu$ and covariance matrices $\mathbf{v}_\sigma$ [29]:

$$\mathbf{v}_\sigma = \left[\mathbf{I} + \sum_c \theta_c \mathbf{T}_c^{\mathrm{T}} \overline{\mathbf{\Sigma}}_c^{-1} \mathbf{T}_c \right]^{-1} \tag{6}$$

$$\mathbf{v}_\mu = \mathbf{v}_\sigma \sum_c \left[\mathbf{T}_c^{\mathrm{T}} \overline{\mathbf{\Sigma}}_c^{-1} \sum_t \theta_{c,t} (\mathbf{o}_t - \mathbf{m}_c) \right] \tag{7}$$

where **I** denotes an identity matrix of appropriate size, $\mathbf{m}_c$ and $\overline{\mathbf{\Sigma}}_c$ are adapted mean and covariance of the c^{th} Gaussian, which are updated during each EM iteration starting from UBM parameters, and $^{\mathrm{T}}$ represents the transpose operator.

In the M-step, the subspace matrix **T** is estimated via maximization of the following auxiliary function over **T**.

$$\widetilde{\Psi}(\gamma, \mathbf{T}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{c=1}^{C} \theta_{c,t,i} \log \pi_{c,i} p(\mathbf{o}_{t,i} \mid [\mu_c + \mathbf{T}_c \mathbf{v}_i], \mathbf{\Sigma}_{c,i}) \tag{8}$$

An efficient procedure for training **T** and for MAP adaptation of the i-vectors can be found in [29]. In the total variability modeling approach, the i-vector is the low-dimensional representation of an audio recording that can be used for classification and estimation purposes.

3) The Non-negative Factor Analysis (NFA) Framework:

The NFA is a new framework for adaptation and decomposition of GMM weights based on a constrained factor analysis [21]. The basic assumption of this method is that for a given utterance, the adapted GMM weight supervector can be decomposed as follows:

$$\mathbf{w} = \boldsymbol{\pi} + \mathbf{L}\mathbf{r}, \tag{9}$$

where $\boldsymbol{\pi}$ is the UBM weight supervector (2048 dimensional vector in this study). $\mathbf{L}$ is a matrix of dimension $C\times\rho$ spanning a low-dimensional subspace (300 dimensions in this work). $\mathbf{r}$ is a low-dimensional subspace vector obtained through a constrained maximum likelihood estimation criterion.

In this framework, the adapted weights are obtained by maximizing the following objective function over w_c .

$$\Psi(\gamma,\mathbf{r}) = \sum_{t=1}^{T}\sum_{c=1}^{C}\theta_{c,t}\log w_c p(\mathbf{o}_t \mid \mu_c, \boldsymbol{\Sigma}_c) \tag{10}$$

Substituting w_c by (π_c+$\mathbf{L}_c\mathbf{r}$) in the Eq. 10, and given an utterance O, a maximum likelihood estimation of $\mathbf{r}$ is obtained by solving the following constrained optimization problem:

$$\max_{\mathbf{r}}\left(\Psi(\gamma,\mathbf{r})\right) = \max_{\mathbf{r}}\left(\bar{\theta}^{\mathrm{T}}(O)\log(\boldsymbol{\pi}+\mathbf{L}\mathbf{r})\right) \tag{11}$$

$$\text{Subject to } \begin{cases}\mathbf{1}(\boldsymbol{\pi}+\mathbf{L}\mathbf{r}) = 1\\ \boldsymbol{\pi}+\mathbf{L}\mathbf{r} > 0\end{cases}$$

where $\mathbf{1}$ is a row vector of dimension C with all elements equal to one, and $\bar{\theta}(O) = \sum_t [\theta_{1,t} \quad \ldots \quad \theta_{C,t}]^{\mathrm{T}}$.

In this framework, neither the subspace matrix $\mathbf{L}$ nor the subspace vector $\mathbf{r}$ is constrained to be non-negative. However, unlike the i-vector framework, the applied factor analysis for estimating the subspace matrix $\mathbf{L}$ and the subspace vector $\mathbf{r}$ is constrained such that the adapted GMM weights are non-negative and sum up to one. The procedure of calculating $\mathbf{L}$ and $\mathbf{r}$ involves a two-stage algorithm similar to EM and can be found in [21]. The subspace matrix $\mathbf{L}$ is estimated over a large training dataset. It is then used to extract a subspace vector $\mathbf{r}$ for each utterance in train and test datasets.

This new low-dimensional utterance representation approach was successfully applied to speaker characterization [23], [24] and language/dialect recognition [21] tasks.

4) Feature-Level Fusion of the i-vectors and NFA Vectors:

Previous studies show that although GMM weight supervectors contain less information than GMM means, they provide complementary information to GMM means [22]. Feature-level fusion and score-level fusion are considered as effective approaches to exploit available information in both GMM means and weights [22], [24]. Score-level fusion, in which the outputs of different estimators are fused, requires a development data set to train the fusion model, which results in decreasing the number of training data. However, fusion at feature level, in which various features are normalized and concatenated, eliminates the need for assigning a considerable amount of available training data for development set, and estimation can be performed in one learning phase.

In this paper, a feature-level fusion of the i-vectors and the NFA vectors is considered to improve the estimation accuracy. As illustrated in Fig. 1, the extracted i-vectors and the NFA vectors are length normalized by having mapped onto a low-dimensional space using linear discriminant analysis (LDA) [30]. Then, the obtained low-dimensional vectors are concatenated to form a longer vector.

B. Function Approximation Using LSSVR

Support vector regression (SVR) is a function approximation approach developed as a regression version of the widely known Support Vector Machines (SVM) classifier. Using nonlinear transformations, SVMs map the input data onto 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 classes, SVR carries out the regression task by finding the optimal regression hyperplane in which most of training samples lie within an ε-margin around this hyperplane [31].

In this study, we use the least squares version of support vector regression. While an SVR solves a quadratic programming with linear inequality constraints, which results in high algorithmic complexity and memory requirement, an LSSVR involves solving a set of linear equations by considering equality constraints instead of inequalities for classical SVR [31], which speeds up the calculations.

In a regression problem, we are given a training dataset $D^{tr} = \{(\mathbf{o}_1, y_1),\ldots,(\mathbf{o}_n, y_n),\ldots,(\mathbf{o}_N, y_N)\}$, where $\boldsymbol{o}_n$ and y_n denote a vector of observed features of the n^{th} data item and its corresponding output, respectively. The goal is to determine a function $h(\boldsymbol{o})$ such that the outputs are predicted accurately. In primal form of LSSVR, $h(\boldsymbol{o})$ is considered as

$$h(\mathbf{o}) = \vartheta^{\mathrm{T}}\varphi(\mathbf{o}) + c \tag{12}$$

A least squares loss function is applied instead of Vapnik's ε-insensitive loss function in LSSVR to simplify the formulations to minimize

$$\frac{1}{2}\|\vartheta\|^2 + \frac{1}{2}\beta\sum_{n=1}^{N} e_n^2 \tag{13}$$

subject to

$$y_n = \vartheta^{\mathrm{T}}\varphi(\mathbf{o}) + c + e_n \tag{14}$$

where β is an error cost factor and $e_n \in$ are error variables.

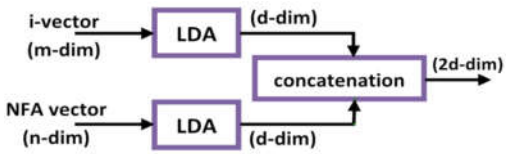

Fig. 1. Block diagram of the utterance modeling in feature-level fusion.

This optimization problem can be solved more efficiently for high dimensional data by using the Lagrangian variables v and minimizing the following dual cost function [31].

$$\Omega(\vartheta,c,e,v)=\frac{1}{2}\|\vartheta\|^2+\frac{1}{2}\beta\sum_{n=1}^{N}e_n^2 - \sum_{n=1}^{N}v_n[\vartheta^{\mathrm{T}}\varphi(\mathbf{o}_n)+c+e_n-y_n] \quad (15)$$

This minimization problem can be solved directly by taking the partial derivative of Ω with respect to ϑ, c, e and v and setting the results to zero. This results in solving a linear system of equations. Inserting the obtained results to (12) leads to the regression function

$$h(\mathbf{o})=\sum_{n=1}^{N}v_n\langle\phi(\mathbf{o}_n),\phi(\mathbf{o})\rangle+c=\sum_{n=1}^{N}v_nK(\mathbf{o}_n,\mathbf{o})+c \quad (16)$$

where $K(\mathbf{o}_n, \mathbf{o})$ denotes the kernel function and c and v are the solution to optimization problem (15).

One drawback of the applied simplification in LSSVR formulation is the loss of sparseness. Therefore, all samples contribute to the model, and consequently, the model often becomes unnecessarily large.

C. Training and Testing

The block diagram of the proposed weight estimation approach is shown in Fig. 2. During the training phase, each utterance in the training data set is mapped to a high dimensional vector using the feature-level fusion utterance modeling described in Section 3.A.4. Then, the obtained vectors along with their corresponding weight labels are used to train an estimator for approximating function g. During the testing phase, the same utterance modeling approach applied in training phase is used to extract a high dimensional vector from a test utterance. Then, the estimated weight is obtained using the trained estimator.

4. Experimental Setup

A. Database

The National Institute for Standard and Technology (NIST) has 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 is released. Conversations typically last 5 minutes and originate from a large number of speakers for whom additional meta-data is recorded.

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 automatic speaker weight estimation, telephone recordings from the common protocols of the recent NIST 2008 and 2010 SRE databases are pooled together to create a dataset of 8241 utterances uttered by 1333 speakers. Then, it is divided to two disjoint parts such that 80% and 20% of all speakers are used for training and testing sets, respectively. Thus, of all 8241 utterances, 5902 utterances are considered for training set and 2339 utterances are considered for testing set. Fig. 3 shows the weight histograms of training and testing datasets for male and female speakers.

B. Performance Metric

In order to evaluate the effectiveness of the proposed method, the mean-absolute-error (MAE) of the estimated weight and the Pearson correlation coefficient (*CC*) between the actual and estimated weights are used. MAE is defined as:

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i-y_i| \quad (17)$$

where $\hat{y}_i$ is the i^{th} estimated weight, y_i is the i^{th} actual weight, and N is the total number of test samples.

The Pearson correlation coefficient is computed as:

$$CC=\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\hat{y}_i-\mu_{\hat{y}}}{\sigma_{\hat{y}}}\right)\left(\frac{y_i-\mu_y}{\sigma_y}\right) \quad (18)$$

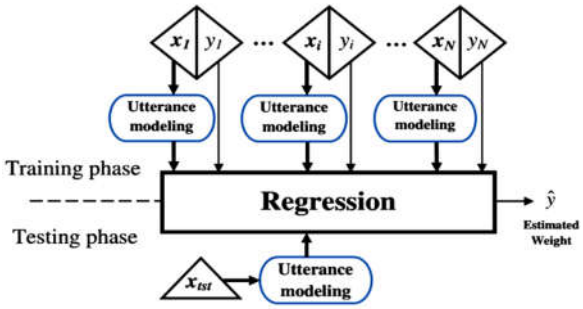

Fig. 2. Block diagram of the proposed speaker weight estimation approach in training and testing phases.

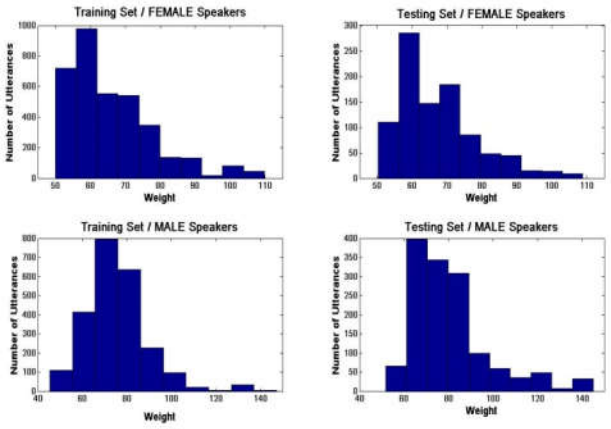

Fig. 3. The weight histograms of telephone speech utterances of training and testing datasets for male and female speakers.

where μ_y and σ_y denote the mean and standard deviation of the actual speakers' weight respectively, and $\mu_{\hat{y}}$ and $\sigma_{\hat{y}}$ are respectively the mean and standard deviation of the estimated weights.

5. Results and Discussion

In this section, the evaluation results of the baseline systems as well as the proposed speaker weight estimation approach are presented. The acoustic feature vector is a 60-dimensional vector consists of 20 Mel-Frequency Cepstrum Coefficients (MFCCs) including energy appended with their first and second order derivatives. This type of feature is very common in the i-vector-based speaker recognition systems. Wiener filtering, feature warping [32] and voice activity detection [33] have also been considered in the front-end processing to obtain more reliable features.

In this study, an LSSVR with a linear kernel has been employed to perform weight estimation, which is implemented using LS-SVMlab1.8 Toolbox [34] in Matlab environment.

A. Results of the Basic Estimation System

When an utterance of an unseen speaker is applied to a basic estimator, its output is the average weight of speakers in training data set. The basic estimation system provides us a chance level accuracy. The results of speaker weight estimation using a basic estimator are reported in the first row of Table I. Besides providing a reference level for speaker weight estimation systems, the basic estimator highlights a limitation of using mean-absolute-error (MAE) as a performance metric for weight estimation problem. The MAE is limited in some respects, specially, in the case of a test set with a skewed distribution which is the case in this task. When a test data set with a skewed distribution is applied to a basic estimator, the MAE might be in an acceptable range, based on the variance of the data. For instance, when the database described in Section IV-A was applied to the basic estimator, the MAE for male and female speakers were 12.93 kg and 9.76 kg, respectively. However, the measured *CC* for males and females were equal to zero. For this reason, the correlation coefficient is a preferred performance metric in this task, which reflects the performance of the estimators in a more sensible way.

B. Results of the i-vector-Based System

In the i-vector-based system, each utterance in training set is mapped onto a 400 dimensional vector using the i-vector framework. Then, the extracted i-vectors along with their corresponding weight labels are used to train estimator. The results of employing an LSSVR as an estimator, and using the i-vector framework for utterance modeling are presented in the second row of Table I. Comparing the i-vector-based system to the basic estimator shows the effectiveness of the i-vectors in automatic speaker weight estimation.

C. The Results of the Score-Level Fusion System

In score-level fusion, in which the outputs of different estimators are fused, we need to allocate a portion of training data for development set to train the fusion model. In this study, the development set consists of utterances of 20% of speakers in training set. In this approach, as illustrated in Fig. 4, each utterance of train, development and test sets are converted to 400 dimensional i-vectors and 300 dimensional NFA vectors. Then, the obtained i-vectors and NFA vectors of the train set are employed to train the i-vector-based and the NFA-based models, respectively. In the next step, the i-vectors and the NFA vectors of development set are applied to the trained models. The outputs are then concatenated to form 2-dimensional vectors and along with the corresponding weight labels are used to train the fusion model. The fusion model is a single hidden layer feedforward neural network with 5 hidden units. Logistic and linear activation functions are considered for the hidden and output neurons, respectively. The network is trained using the one step secant back-propagation algorithm [35] which is implemented using Neural Network Toolbox [36] in Matlab environment.

The results of the proposed score-level fusion system for speaker weight estimation are presented in the third row of Table 1. Comparing to the results of the i-vector-based estimator, it can be concluded that fusion of the outputs of two subsystems can slightly improve the estimation accuracy which indicates that GMM weights carry complementary information to GMM means. The achieved relative improvements in *CC* by the proposed fusion scheme compared to the i-vector-based estimator for male and female speakers are 2.32% and 6.25%, respectively.

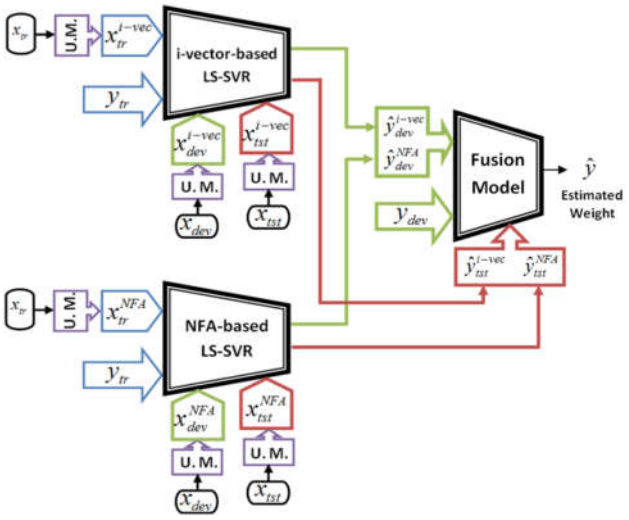

Fig. 4. Block diagram of the score-level fusion speaker weight estimation system (U. M. stands for utterance modeling).

D. Results of the Proposed Approach

To improve the estimation accuracy of the i-vector-based weight estimation, a feature-level fusion of the i-vectors and the NFA vectors is considered in this paper. In the proposed method, the extracted i-vectors and NFA vectors are length normalized and concatenated to form a longer vector. The obtained vector, along with the corresponding weight label is then used to train estimator. The last row of Table 1 contains the results of the proposed weight estimation approach using a

fusion of the i-vectors and the NFA vectors. The obtained results indicate that the accuracy of weight estimation increases after feature-level fusion compared to the estimation using the i-vector-based estimator, which again shows that GMM weights provide complementary information to GMM means.

The achieved relative improvements in *CC* by the proposed feature-level fusion scheme compared to the i-vector-based estimator for male and female speakers are 25% and 38.77%, respectively. Comparing the results of these two fusion schemes reveals that fusion of the i-vector and the NFA frameworks at feature level is more effective in speaker weight estimation. In addition, fusion at feature level eliminates the need for assigning a considerable amount of training data for development set, and speaker weight estimation is performed in one learning phase.

The reported *CC* for speaker weight estimation based on the formant parameters of the running speech signals uttered by 91 speakers are 0.33 and 0.34 for male and female speakers, respectively [9]. The results obtained from our proposed speaker weight estimation system seem reasonable, considering the fact that the applied testing dataset in this study consists of spontaneous speech signals and the number of speakers in test set is considerably larger than that of in [9]. It can be concluded that automatic speaker weight estimation using a fusion of the i-vector and the NFA frameworks is more efficient compared to the estimation based on the raw acoustic features.

Table 1. The MAE (in kg) and CC of the proposed speaker weight estimation systems, compared to the basic and i-vector-based systems.

Speaker Weight Estimation System	**MALE**		**FEMALE**	
	CC	**MAE**	***CC***	**MAE**
Basic estimator	0	12.93	0	9.76
i-vector-based system	0.42	12.17	0.30	9.70
Score-level fusion subsystems	0.43	11.98	0.32	9.30
Feature-level fusion system	**0.56**	**11.16**	**0.49**	**7.79**

6. Conclusion

In this paper a novel approach for automatic speaker weight estimation from spontaneous telephone speech signals was proposed. In this method, each utterance was modeled using a fusion of the i-vector and the NFA frameworks at feature level. Using this new utterance modeling approach, the available information in both GMM means and GMM weights was utilized. Then, an LSSVR was employed to estimate the weight of a speaker from a given utterance. The proposed method was trained and tested on the telephone conversations of NIST 2008 and 2010 SRE corpora. Evaluation results over 2339 utterances show that the correlation coefficients between the actual and the estimated weights of male and female speakers after feature-level fusion are 0.56 and 0.49, respectively, which indicate that the fusion of the i-vectors and the NFA vectors at feature level improves the performance of the state-of-the-art i-vector framework. Utilizing information in Gaussian weights in conjunction with that of in Gaussian means through a fusion of the i-vector and the NFA frameworks resulted in achieving 25% and 38.77% relative improvements in CC compared to the i-vector-based weight estimation system. It also indicates the effectiveness of the proposed method in automatic speaker weight estimation compared to the estimation based on the raw acoustic features.

REFERENCES

[1] M. H. Bahari, M. McLaren, H. Van hamme, and D. Van Leeuwen, "Age estimation from telephone speech using i-vectors," *Eng. Appl. Artif. Intell. Elsevier*, vol. 34, pp. 99–108, 2014.

[2] A. H. Poorjam, M. H. Bahari, and H. Van Hamme, "Speaker weight estimation from speech signals using a fusion of the i-vector and NFA frameworks," in *International Symposium on Artificial Intelligence and Signal Processing (AISP)*, 2015, pp. 118–123.

[3] J. Harrington and S. Cassidy, "The Acoustic Theory of Speech Production," in *Techniques in Speech Acoustics*, Dordrecht: Springer Netherlands, 1999, pp. 29–56.

[4] N. J. Lass and M. Davis, "An investigation of speaker height and weight identification.," *J. Acoust. Soc. Am.*, vol. 60, no. 3, pp. 700–3, Sep. 1976.

[5] C. Darwin, *The Descent of Man and Selection in Relationship to Sex*. Princeton University Press, 1981.

[6] N. J. Lass, "Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies.," *J. Acoust. Soc. Am.*, vol. 63, no. 4, pp. 1218–20, Apr. 1978.

[7] H. J. Künzel, "How well does average fundamental frequency correlate with speaker height and weight?," *Phonetica*, vol. 46, no. 1–3, pp. 117–25, 1989.

[8] W. T. Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques.," *J. Acoust. Soc. Am.*, vol. 102, no. 2 Pt 1, pp. 1213–22, Aug. 1997.

[9] J. González, "Formant frequencies and body size of speaker: a weak relationship in adult humans," *Journal of Phonetics*, vol. 32, no. 2. pp. 277–287, 2004.

[10] U. G. Goldstein, "An articulatory model for the vocal tracts of growing children," Massachusetts Institute of Technology, 1980.

[11] V. E. Negus, "The comparative anatomy and physiology of the larynx," *Laryngoscope*, vol. 60, no. 5, pp. 516–516, May 1950.

[12] W. A. van Dommelen and B. H. Moxness, "Acoustic parameters in speaker height and weight identification: sex-specific behaviour.," *Lang. Speech*, vol. 38 (Pt 3), pp. 267–87.

[13] W. M. Campbell, D. E. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[14] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[15] M. H. Bahari, "Automatic Speaker Characterization: Automatic Identification of Gender, Age, Language and Accent from Speech Signals," PhD Thesis, University of Leuven (KU Leuven), 2014.

[16] A. H. Poorjam, M. H. Bahari, V. Vasilakakis, and H. Van hamme, "Height estimation from speech signals using i-vectors and least-squares support vector regression," in *37th Intenational Conference on Telecommunications and Signal Processing*, 2014, pp. 1–5.

[17] A. H. Poorjam, R. Saeidi, T. Kinnunen, and V. Hautam, "Incorporating uncertainty as a quality measure in i-vector based language recognition," in *Odyssey*, 2016, pp. 74–80.

[18] K. A. Lee and et. al, "The 2015 NIST Language Recognition Evaluation : the Shared View of I2R, Fantastic4 and SingaMS," in *ICASSP*, 2016, pp. 3211–3215.

[19] M. H. Bahari and H. Van hamme, "Speaker age estimation using Hidden Markov Model weight supervectors," in *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, 2012, pp. 517–521.

[20] M. H. Bahari and H. Van Hamme, "Speaker age estimation and gender detection based on supervised Non-Negative Matrix Factorization," in *2011 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, 2011, pp. 1–6.

[21] M. H. Bahari, N. Dehak, H. Van hamme, L. Burget, A. M. Ali, and J. Glass, "Non-Negative Factor Analysis of Gaussian Mixture Model Weight Adaptation for Language and Dialect Recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 7, pp. 1117–1129, 2014.

[22] 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 *International conference on acoustics, speech, and signal processing (ICASSP)*, 2013, pp. 7344–7348.

[23] A. H. Poorjam, "Speaker Profiling for Forensic Applications," Master's Thesis, University of Leuven (KU Leuven), 2014.

[24] A. H. Poorjam, M. H. Bahari, and H. Van hamme, "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *Int. Conf. on Computer and Knowledge Engineering*, 2014, pp. 7–12.

[25] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.

[26] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification," in *Odyssey*, 2010.

[27] N. Dehak, "Discriminative and generative approaches for long- and short-term speaker characteristics modeling application to speaker verification.," École de technologie supérieure, 2009.

[28] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-Class Covariance Normalization for SVM-based Speaker Recognition," in *INTERSPEECH* , 2006.

[29] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[30] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2000.

[31] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. WORLD SCIENTIFIC, 2002.

[32] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *Odyssey*, 2001.

[33] M. McLaren and D. A. Van Leeuwen, "A simple and effective speech activity detection," in *NIST SRE 2011 workshop*, 2011.

[34] K. De Brabanter, P. Karsmakers, O. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, and J. A. K. Suykens, "Ls-svmlab1.8 toolbox." http://www.esat.kuleuven.be/sista/lssvmlab/.

[35] R. Battiti and Roberto, "First- and Second-Order Methods for Learning: Between Steepest Descent and Newton's Method," *Neural Comput.*, vol. 4, no. 2, pp. 141–166, Mar. 1992.

[36] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*. Boston: PWS Publishing Co., 1997.

Amir Hossein Poorjam was born in Esfahan, Iran, in June 1986. He received his M.Sc. degree in Electrical Engineering from University of Leuven (KU Leuven), Belgium, in 2014. During fall and winter 2015, he visited the Speech and Image Processing Unit (SIPU), at University of Eastern Finland, where he was granted a CIMO fellowship. He is currently pursuing the Ph.D. degree in the Audio Analysis Laboratory at Aalborg University, Denmark. His research interests include speech signal processing, speaker profiling and language recognition.

Mohamad Hasan Bahari is a postdoctoral researcher at STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven. He received his Ph.D. from KU Leuven, Belgium, where he was granted a Marie-Curie fellowship. During winter, spring and fall 2013, he visited the Computer Science and Arti cial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT). His works on Data Analytics and machine learning was awarded the International Speech Communication Association (ISCA) best student paper at INTERSPEECH 2012 and second place of Reinforcement Learning Competition (Octopus event) 2009.

Hugo Van hamme received the Master's degree in engineering (burgerlijk ingenieur) from VUB in 1987, the M.Sc. degree from Imperial College, London in 1988 and the Ph.D. degree in Electrical Engineering from Vrije Universiteit Brussel (VUB) in 1992. From 1993 till 2002, he worked for L&H Speech Products and ScanSoft, initially as senior researcher and later as research manager. Since 2002, he is a professor at the department of electrical engineering of KU Leuven. His main research interests are: applications of speech technology in education and speech therapy, computational models for speech recognition and language acquisition and noise robust speech recognition.