Age Estimation from Telephone Speech using i-vectors

Mohamad Hasan Bahari[1], *Mitchell McLaren*[2], *Hugo Van hamme*[1], *David Van Leeuwen*[2]

[1]Center for processing speech and images, KU Leuven, Belgium
[2]Center for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

MohamadHasan.Bahari@esat.kuleuven.be, m.mclaren@let.ru.nl,
hugo.vanhamme@esat.kuleuven.be, d.vanleeuwen@let.ru.nl

Abstract

Motivated by the success of i-vectors in the field of speaker recognition, this paper proposes a new approach for age estimation from telephone speech patterns based on i-vectors. In this method, each utterance is modeled by its corresponding i-vector. Then, Support Vector Regression (SVR) is applied to estimate the age of speakers. The proposed method is trained and tested on telephone conversations of the National Institute for Standard in Technology (NIST) 2010 and 2008 Speaker Recognition Evaluations databases. Evaluation results show that the proposed method outperforms different conventional methods in speaker age estimation.

Index Terms: speaker age estimation, i-vector, support vector regression

1. Introduction

Speaker age estimation has recently received increased attention due to its wide range of commercial applications such as interactive voice response systems, targeted advertising, and service customization [1, 2, 3]. It can also be applied in forensic scenarios to narrow down the number of suspects. However, automated speech-based age estimation is challenging for a number of reasons. First, there usually exists a difference between the perceived age of speakers and their actual age (or chronological age). Second, developing a robust age estimation method requires a database of speech from age-labeled speakers with a wide, yet balanced, range of ages. Third, speech contains significant intra-speaker variability that is not related to, or closely correlated with, age [4] including speaker weight, height and emotional condition. One effective approach to age estimation from speech involves modeling speech recordings with Gaussian Mixture Model (GMM) mean supervectors before use as features in Support Vector Regression (SVR) [1, 2].

Similar Support Vector Machine (SVM) techniques have been successfully applied to different speech analysis problems such as speaker recognition [5]. While effective, GMM mean supervectors are of a high dimensionality resulting in high computational cost and difficulty in obtaining a robust model in the context of limited data. Consequently, dimension reduction through PCA-based methods has been found to improve performance of age estimation from GMM mean supervectors [1].

In the field of speaker recognition, recent advances using so-called i-vectors [6] 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. The same idea was also applied in speaker language recognition effectively [7]. In this paper, we replace GMM mean supervectors by low-dimensional i-vectors to model utterances in a speaker age estimation system. Evaluation on the NIST 2010 and 2008 SRE databases shows that the accuracy of the proposed speaker age estimator increases due to more sophisticated representation of speech patterns.

The rest of this paper is organized as follows. In Section 2 the problem of speaker age estimation and different conventional approaches addressing this issue are described. In section 3, the proposed approach is elaborated. Section 4 explains our experimental setup. The evaluation results are presented and discussed in section 5. The paper ends with conclusions in section 6.

2. Age Estimation from Speech

In speaker age estimation, we are given a training dataset of speech recordings $S^{tr} = \{(x_1, y_1), \ldots, (x_n, y_n), \ldots, (x_N, y_N)\}$. In this set x_n and y_n denote the n^{th} utterance of the training dataset and its corresponding speaker age, respectively. The goal is to design an estimator function g, such that for an utterance of an unseen speaker x^{tst}, the difference between estimated age $\hat{y} = g(x^{tst})$ and actual age is minimized.

2.1. Baseline Approaches

Prior: The most basic choice for the estimator function is the average age of the training data, $g(x^{\text{tst}}) = \frac{1}{N}\sum_n y_n$. This estimator, labeled as prior in the rest of this paper, intuitively provides a low level of accuracy.

GMM-SVR: Different methods have been introduced to reach an effective speaker age estimation [1]–[4]. For example, Bocklet et al. introduced GMM-SVR to estimate the age of children from GMM mean supervectors derived from their voice patterns [2]. Given an utterance, Maximum A Posteriori adaptation (MAP) is applied to adapt a Universal Background Model (UBM) to the speech characteristics of the speaker. Component means of the obtained GMM are then extracted and concatenated to form a GMM mean supervector representing the utterance. Finally, SVR as a function approximator is applied to estimate the speakers' age.

GMM-PCA-SVR and **GMM-WPPCA-SVR:** The approach of GMM-SVR was adopted and extended by Dobry and his colleagues [1] by applying dimension reduction techniques to the superverctor. Methods such as Principle Component Analysis (PCA) and Weighted-Pairwise PCA (WPPCA) were applied and investigated. It was concluded that WPPCA, which is a supervised dimensionality reduction approach working based on nuisance attribute projection, yields more accurate results. These speaker age estimators, labeled GMM-PCA-SVR and GMM-WPPCA-SVR, are used as contrastive baseline systems

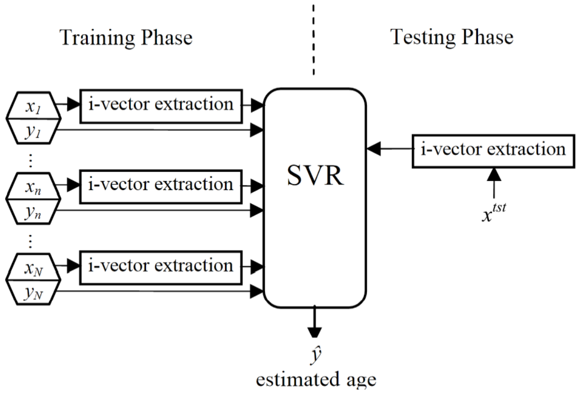

Figure 1: *The block diagram of the proposed speaker age estimation approach in training and testing phases.*

in this paper.

3. Age Estimation using i-vectors

The approaches to age estimation described in section 2.1 are based on GMM mean supervectors and have been shown to provide a good level of performance. In the related field of speaker recognition, GMM supervectors are commonplace. Recent progress in this field, however, has found an alternate method of modeling GMM supervectors that provides far superior speaker recognition performance [6]. This technique is referred to as total variability modeling. Total variability modeling assumes the GMM mean supervector, $\boldsymbol{\mu}$, that best represents a set of feature vectors can be decomposed as

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

where $\mathbf{m}$ is the mean supervector of the UBM, $\mathbf{T}$ spans a low-dimensional total variability subspace (400 dimensions in this work) and $\mathbf{w}$ are the factors that best describe the utterance-dependent mean offset $\mathbf{T}\mathbf{w}$. The vector $\mathbf{w}$ is commonly referred to as the i-vector and has a standard normal distribution $N(0, I)$. Subspace $\mathbf{T}$ is estimated via factor analysis to represent the directions that best separate different speech recordings in a large development dataset. An efficient procedure for training $\mathbf{T}$ and MAP adaptation of i-vectors $\mathbf{w}$ can be found in [8]. In the total variability modeling approach, i-vectors are the low-dimensional representation of an audio recording that can be used for classification and estimation purposes. For the purpose of age estimation, we replace GMM supervectors with corresponding i-vectors in GMM-SVR system described in section 2.1.

The principle of the proposed age estimation approach is illustrated in figure 1. The applied SVR uses a Gaussian kernel function and 5-fold cross-validation to tune the smoothing parameter of the kernels.

The use of i-vectors for age estimation has several distinct advantages over GMM supervectors. Firstly, the relatively low dimensionality of i-vectors (400) significantly reduces the computational burden of model training and estimation compared to a GMM supervector dimensionality of greater than 12,000 used in this work. Secondly, subspace adaptation of i-vector $\mathbf{w}$ results in a more reliable estimation of true model means $\boldsymbol{\mu}$ in the context of limited training data.

3.1. I-vector Session Compensation

Perhaps one of the most dominant topics in speaker recognition is that of session compensation. Session compensation aims to remove session variation or within-class variation from feature vectors (such as GMM supervectors or i-vectors) to allow the subsequent modeling techniques to better observe important between-class information. In the context of age estimation, session variation is anything that makes features corresponding to the same age appear different and encompasses microphone types, transmission channels, language, gender and even the speaker. As i-vectors are inherently good at representing speaker characteristics along with other sources of session variation, we aim to compensate for this variation using two widely accepted techniques for speaker recognition using i-vectors—Linear Discriminant Analysis (LDA) and Within-class Covariance Normalization (WCCN). Since age is a continuous variable, we have to categorize speakers into G age groups to apply LDA OR WCCN.

3.1.1. LDA

LDA is a dimensionality reduction method which transforms the feature space such that the between-class scatter S_b is maximized and simultaneously the within-class scatter S_w is minimized [9]. This is obtained through the eigenvalue decomposition of relation 2. The LDA projection matrix, A, is formed as the subset of eigenvectors, θ, having the largest eigenvalues, λ.

$$S_b\theta = \lambda S_w\theta \tag{2}$$

$$S_b = \sum_{g=1}^{G} N_g \left(\bar{w}_g - \bar{w}\right)\left(\bar{w}_g - \bar{w}\right)^t \tag{3}$$

$$S_w = \sum_{g=1}^{G}\sum_{i=1}^{N_g} \left(w_g^i - \bar{w}_g\right)\left(w_g^i - \bar{w}_g\right)^t \tag{4}$$

where $\bar{w}_g = \frac{1}{N_g}\sum_i^{N_g} w_g^i$ is the mean of the observations in the g^{th} age category, $\bar{w}$ represents the mean of all instances in the training set, w_g^i is the i^{th} i-vector in the g^{th} age category, N_g denotes the number of utterances in the g^{th} age category and t is means vector transpose.

3.1.2. WCCN

Within-Class Covariance Normalization (WCCN) aims to normalize the within-class covariance of the i-vector space to the identity matrix. In doing so, directions of relatively high within-class variation will be attenuated and thus prevented from dominating the space [10]. We investigate both methods in section 5.3. The WCCN transformation matrix B is found through the Cholesky decomposition of

$$\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{N_g}\sum_{i=1}^{N_g}\left(w_g^i - \bar{w}_g\right)\left(w_g^i - \bar{w}_g\right)^t\right]^{-1} = BB^t \tag{5}$$

4. Experimental Setup

4.1. Database

The National Institute for Standard in Technology (NIST) have held annual or biannual Speaker Recognition Evaluations (SRE) for the past two decades. With each SRE, a large database of telephone (and more recently microphone) conversations are released along with an evaluation protocol. These

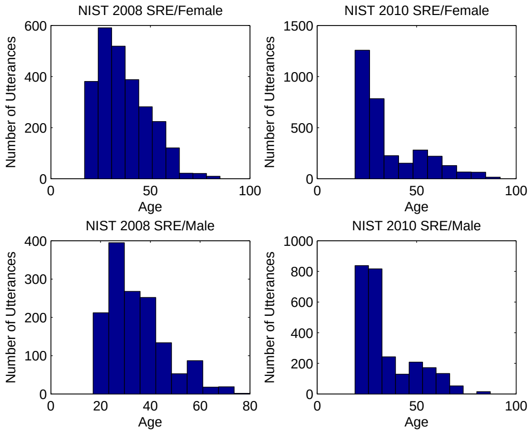

Figure 2: *Age histogram of telephone speech utterances for NIST 2010 and 2008 SRE Databases.*

conversations typically last 5 minutes and originate from a large number of participants for whom meta data is recorded—including participant age. The NIST databases where 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 dataset used to train the total variability subspace and UBM includes over 30,000 speech recordings and was sourced from NIST 2004–2006 SRE databases, LDC releases of Switchboard 2 phase III and Switchboard Cellular (parts 1 and 2). For the purpose of age estimation, telephone recordings from the common protocols of the recent NIST 2010 and 2008 SRE databases are used for training and testing respectively. The core protocol, short2-short3, from the 2008 database contains 3999 telephone recordings for 1336 speakers for whom the age is known. Similarly, the extended core-core protocol of the 2010 database contains 5634 telephone speech segments from 445 speakers. Figure 2 illustrates the age histograms of male and female speakers of NIST 2010 and 2008 SRE databases.

The effectiveness of the proposed method is evaluated using the Mean Absolute Error (MAE) of the estimated speakers age, which is calculated as follows.

$$\mathrm{MAE} = \frac{1}{Q}\sum_{q=1}^{Q} |\hat{y}_q - y_q| \tag{6}$$

where $\hat{y}_q$ and y_q are the estimated and the chronological age of the q^{th} utterance of the testing dataset respectively. Q is the total number of utterances in the testing dataset .

5. Results and Discussion

This section provides the evaluation results of the baseline systems and compares them to the proposed method using i-vectors. In this problem, speakers from the same birth year belong to the same class.

5.1. Baseline Systems Results

In this section, the performances of baseline systems, namely prior, GMM-SVR, GMM-PCA-SVR and GMM-WPPCA-SVR, are investigated.

The applied GMM in all baseline systems consist of 512 mixture components. To study the effect of the acoustic features, two types of feature vectors have been tested for the baseline systems. The first type, labeled $\mathrm{MFCC_{26D}}$, consists of 13 Mel-Frequency Cepstrum Coefficients (MFCCs) including energy appended with their first order derivatives, forming a 26 dimensional acoustic feature vector. The second type, $\mathrm{MFCC_{60D}}$, consists of 20 MFCCs including energy appended with their first and second order derivatives, forming a 60 dimensional acoustic feature vector.

The former type, $\mathrm{MFCC_{26D}}$, matches the configuration of features applied in [4] and the latter type, $\mathrm{MFCC_{60D}}$, is very common in state-of-the-art i-vector based speaker recognition systems. To have clean features, speech activity detection, feature warping and Wiener filtering have also been applied in the feature extraction phase. The MAE of male and female speakers' age estimation using the baseline systems with both types of acoustic features are listed in table 1. In this experiment, PCA and WPPCA have been tested over different target dimensions between 100 and 1000. Table 1 only includes the best results, which were obtained on target dimensions 200 and 300 for PCA and WPPCA respectively.

Results in table 1 indicate that the GMM-SVR system is remarkably more accurate than the Prior system. This shows that the GMM supervectors contain speaker information including age. The table also shows that the PCA and WPPCA-based systems outperformed the GMM-SVR system, thus demonstrating the benefit of dimension reduction of the GMM supervectors prior to SVR. Unlike [4], our experiments do not show any advantage for using WPPCA over PCA. It is also interpreted from table 1 that increasing the acoustic dimension from 26 to 60 improves the estimation accuracy. Therefore, in the rest of our experiments we focused on the second type of acoustic features, $\mathrm{MFCC_{60D}}$.

5.2. I-vectors for Age Estimation

The results of applying the proposed method for speakers' age estimation are presented in this section.

Figure 3 presents the MAE of the estimated age using the proposed method and baseline systems in different target dimensions for female and male speakers respectively. This figure show that i-vector-SVR is more accurate than the other state-of-the-art approaches. The proposed approach improves the age

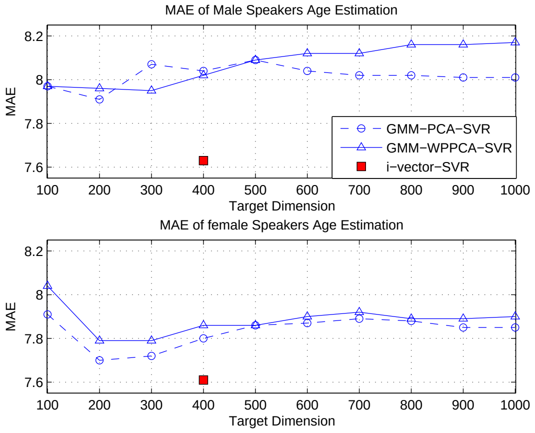

Figure 3: *The MAE of female and male speakers' age estimation using the proposed method and baseline systems versus target dimension.*

Table 1: *The MAE (in years) of male and female speakers' age estimation for the baseline systems using* $\mathrm{MFCC_{26D}}$ *and* $\mathrm{MFCC_{60D}}$ *feature vectors.*

System Configuration	Female		Male	
	$\mathrm{MFCC_{26D}}$	$\mathrm{MFCC_{60D}}$	$\mathrm{MFCC_{26D}}$	$\mathrm{MFCC_{60D}}$
Prior	9.34	9.34	10.39	10.39
GMM-SVR	8.63	8.24	8.16	8.02
GMM-PCA-SVR	7.91	7.91	7.94	7.7
GMM-WPPCA-SVR	7.88	7.95	8.1	7.79

estimation accuracy of all 3999 test set utterances by 5.94 %, 2.93 %, and 2 % relative to GMM-SVR, GMM-WPPCA-SVR, and GMM-PCA-SVR respectively. Note that this improvement was obtained by simply using i-vectors instead of GMM mean supervectors in GMM-SVR. Consequently, in i-vector-SVR no optimization over target dimension is performed. Therefore, in figure 3, the result of i-vector-SVR is only shown for dimension 400.

5.3. I-vector Session Compensation

The effect of LDA and WCCN on proposed age estimator is investigated in the following experiments.

In first experiment, the training set utterances were labeled by their speakers' birth year. Then, the i-vector of the training set utterances along with their corresponding labels were used to obtain a LDA transformation matrix. The obtained transformation matrix is applied to reduce the dimension of training and testing set i-vectors. Finally, the SVR is employed over the transformed sets to estimate the speakers' age. This experiment was repeated over different target dimensions. Table 2 shows the best results of this method with label i-vector-LDA-SVR, obtained on target dimension 28.

In the second experiment, WCCN uses the i-vectors of the training dataset utterances and their labels to calculate a transformation matrix, which is applied to normalize the within-class covariance of the i-vector space to the identity matrix. Like the last experiment, the SVR is employed over the transformed sets to estimate the speakers' age and the results of this method, namely i-vector-WCCN-SVR, are included in table 2.

As can be seen from table 2, neither i-vector-LDA-SVR nor i-vector-WCCN-SVR are helpful in this problem. The large intra-class variation caused by speaker, channel, etc. differences results in significant overlap of the classes, as can be witnessed by the MAE which spans several age classes. With smaller class differences, more data is required before the discriminant subspace directions emerge from the estimation noise, even more so under a model describing classes with only second order statistics. Apart from a lack of data, LDA and WCCN do not take into account that age is an ordinal number and that hence some class differences should be more important than others. Therefore, further investigations are required to solve the session compensation problem for a continuous property like age.

6. Conclusions

In this paper, utterance modeling with i-vectors, which was successfully applied to speaker recognition, has been used in conjunction with an SVR to address speaker age estimation. To evaluate the proposed estimator, telephone utterances of NIST 2010 and 2008 SRE databases have been used for training and testing respectively. Assessment results confirm the effectiveness of the proposed approach.

Table 2: *The MAE (in years) of male and female speakers' age estimation for i-vector-LDA-SVR, i-vector-WCCN-SVR, and i-vector-SVR.*

	Female	Male
i-vector-LDA-SVR	8.19	8.54
i-vector-WCCN-SVR	8.07	9.01
i-vector-SVR	7.61	7.63

7. Acknowledgements

This work is supported by the European Commission as a Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008), namely Bayesian Biometrics for Forensics (BBfor2), under Grant Agreement number 238803.

8. References

[1] Dobry, G., et al.,"Supervector Dimension Reduction for Efficient Speaker Age Estimation Based on the Acoustic Speech Signal", IEEE Trans. Audio, Speech and Language Processing, 19:1975–1985, 2011.

[2] Bocklet, T., Maier, A. and Noth, E.,"Age Determination of Children in Preschool and Primary School Age with GMM-Based Supervectors and Support Vector Machines/Regression", in Proc. 11th Int. Conf. Text, Speech and Dialogue, Berlin,1:253–260, 2008.

[3] Li, M., et al.,"Automatic speaker age and gender recognition using acoustic and prosodic level information fusion", Comput. Speech Lang., 2012, accepted for publication.

[4] Bahari, M. H., Van hamme, H.,"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, Italy, 27–32, 2011.

[5] Campbell, W. M., Sturim, D. E. and Reynolds, D. A.,"Support vector machines using GMM supervectors for speaker verification", IEEE Signal Processing Letters, 13(5): 308–311, 2006.

[6] Dehak, N., et al.,"Front-end factor analysis for speaker verification", IEEE Trans. Audio, Speech and Language Processing, 19(4): 788–798, 2011.

[7] Dehak, N., et al.,"Language Recognition via Ivectors and Dimensionality Reduction", in Proc. Interspeech 2011, Firenze, Italy, 2011.

[8] Kenny, P., et al.,"A Study of Interspeaker Variability in Speaker Verification", IEEE Trans. Audio, Speech and Language Processing, 16(5): 980–988, July 2008.

[9] Duda, R.O., Hart, P.E., Stork, D.G. "Pattern Classification", Wiley Interscience, 2000.

[10] Hatch, A.O., et al.,"Within-class covariance normalization for SVM-based speaker recognition", in proc. INTERSPEECH, 2006.