SIMPLIFIED AND SUPERVISED I-VECTOR MODELING FOR SPEAKER AGE REGRESSION

Prashanth Gurunath Shivakumar[1], *Ming Li* [23], *Vedant Dhandhania*[1] *and Shrikanth S.Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, USA
[2]SYSU-CMU Joint Institute of Engineering, Sun Yat-Sen University, Guangzhou, China
[3]SYSU-CMU Shunde International Joint Research Institute, Guangdong, China

ABSTRACT

We propose a simplified and supervised i-vector modeling scheme for the speaker age regression task. The supervised i-vector is obtained by concatenating the label vector and the linear regression matrix at the end of the mean super-vector and the i-vector factor loading matrix, respectively. Different label vector designs are proposed to increase the robustness of the supervised i-vector models. Finally, Support Vector Regression (SVR) is deployed to estimate the age of the speakers. The proposed method outperforms the conventional i-vector baseline for speaker age estimation. A relative 2.4% decrease in Mean Absolute Error and 3.33% increase in correlation coefficient is achieved using supervised i-vector modeling using different label designs on the NIST SRE 2008 dataset male part.

Index Terms**—** Supervised i-vector, i-vector, speaker age regression, age recognition, simplified supervised i-vector

1. INTRODUCTION

Speech contains valuable information about the linguistic context as well as speaker identity and paralinguistic information about speaker state and speaker trait. These include information such as the emotional state, gender and age [1]. Age can be an important source of information in many user centered applications. In certain scenarios, speech is the only source of data available. Systems that estimate age from speech utterances can have wide applications. These include automating entries to places which require a minimum or maximum age, targeted advertisement, effective call diverting in call centers, personalized educational systems amongst others. Robust speaker age estimators could also benefit speech recognition problems arising due to speaker age variation [2].

Speaker age regression is a difficult estimation problem for several reasons. First, the large difference between perceptual age (as perceived by humans from speech cues) and chronological age [3] makes it a tough problem as compared to language recognition, and speaker recognition where the references are closely matched with the data characteristics. Second, speaker age is a continuous variable making it difficult to estimate by machine learning methods working with discrete labels. Research groups have treated it either as a classification [4][5] or as a regression problem [6]. Third, very few publicly available age labeled data sets exist which have adequate number of speech utterances from a variety of age groups. Finally, speech contains significant intra-age variability due to identity, speaking style, speech content, emotional states, etc., which makes the speakers of same age sound different [6].

Several systems have been proposed for estimating age using speech [6][7][4][5]. We discuss methods and results that are closely related to ours and those that have motivated the present work. Recently, total variability i-vector modeling with backend variability compensation has gained significant attention in language recognition [8] and speaker verification [9] domains due to its excellent performance, low complexity and compact representation.

Bahari et al. [6] model speaker utterances by well known traditional i-vectors and use Support Vector Regression (SVR) to achieve a best mean absolute error (MAE) of 7.6 years. However, using speaker age as label information for Within Class Covariance Normalization (WCCN) and Linear Discriminant Analysis (LDA) was reported to be not effective which motivates our study.

Bocklet et al. [10] combine age and gender classification as a 7 class age-gender problem using GMM supervectors based Support Vector Machines (SVM) on the SpeechDat 2 corpus. The best overall precision is 77%. The 4 age groups are chosen as Children ($<$ 13 years), Young (14-19 years), Adult (20-64 years), Senior ($>$ 64 years) in their study for both males and females. Three well known kernels were used: polynomial, radial basis function (RBF) and a GMM based distance kernel. In most similar age classification systems the Adult age group has a wide range (almost 45 years) making such classification systems less suited for certain applications. This also serves as a motivation for us to consider the age estimation as a regression problem rather than a classification problem.

In a related work [11], Li et al. have proposed a simplified supervised i-vector model for speaker verification. The traditional i-vector was extended to a label supervised i-vector by concatenating a binary label vector and the linear classifier matrix at the end of the mean supervector and the i-vector factor loading matrix, respectively. The supervised i-vector was shown to be more discriminative.

In this paper, we apply the simplified and supervised i-vector framework [11] to the age regression task. We modify the process of traditional i-vector training by adding the age label information of the training data to form a supervised i-vector framework. We propose novel label vector designs and evaluate their performances under the age regression task on the NIST SRE 2008 dataset.

2. AGE REGRESSION USING DIFFERENT I-VECTOR METHODS

2.1. The i-vector

Fig. 1. depicts the i-vector generation process. A given speech utterance is represented by a new vector called the i-vector x. The advantage of the i-vector is that the high dimensional GMM mean supervector, $\tilde{F}$, obtained by concatenating the 1^{st} order Baum-Welch statistics vector, can be represented by a low dimensional i-vector with the corresponding subspace. The objective behind the process is to project the supervector into a low dimensional subspace such

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

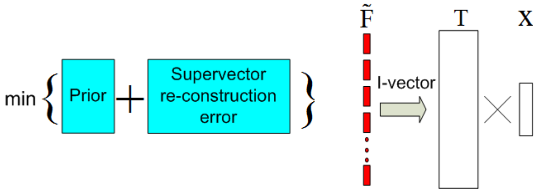

Fig. 1. *The I-vector framework*

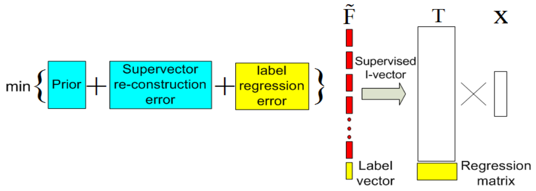

Fig. 2. *The Supervised I-vector framework*

that the sum of the prior and the supervector reconstruction error is minimized.

2.2. The Supervised i-vector

Fig. 2 depicts the framework of the supervised i-vector. The i-vector training is supervised by concatenating the label vector at the end of the GMM mean supervector, $\tilde{F}$. The linear regression matrix is appended at the end of the total variability matrix, T, to reconstruct the label. This supervised i-vector, x, is optimized not only to reconstruct the mean supervectors well but also to predict the age label.

2.3. The Simplified Supervised i-vector

In this work, we adopted the framework in [11] to apply a simplified version of both i-vector and supervised i-vector to reduce the complexity involved.

2.4. Intersession Compensation

In i-vector modeling, the total variability space contains both the speaker and the channel variabilities together, hence session variability compensation needs to be used. In the case of the speaker age estimation problem, session variation makes utterances from the same age class sound different. It may be due to factors like gender, language, transmission channels, microphone types, emotional condition or even the speaking style variability. Variability compensation aims to reduce the within-class variance and allow the modeling technique to observe the inter-class information more effectively. We look at two well known intersession compensation techniques, LDA and WCCN.

LDA is used for dimensionality reduction. It seeks new orthogonal axes to make different classes better discriminated from one another. The axes are such that the inter-class variance is increased and the intra-class variance is decreased [12].

WCCN aims to reduce the within class covariance by normalizing the i-vectors which is typically useful for verification tasks [13].

In our experiments, for both WCCN and LDA, we consider each unique age category in the training dataset as a class.

3. LABEL MODELING TECHNIQUES

For the task of Speaker Age regression, we need to consider each individual age as a class and make each class more separable from the other. Suppose there are M different classes to which speakers belong in the training data then the label matrix L is M x Γ where Γ is the total number of utterances in the training data set. In this section we propose different label vector designs.

3.1. Binary Labels

The default label design in [11] is the binary label where every age label dimension is given a binary value 1 or 0 depending on whether the utterance belongs to this particular class or not. It is defined as follows:

$$L_{ij} = \begin{cases} 1 & \text{if utterance } j \text{ is from age class } i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

If there are M age classes, then L_j is a M dimension binary vector.

3.2. Gaussian Labels

The large difference between the perceived age and the chronological age motivated us to assign soft weights to each class. For example, a speaker of age 28 may sound similar to those in the near vicinity of that age. The class to which the utterance belongs is given the maximum weight and the adjacent classes are given smaller weights. For each particular age value, the corresponding label vector is generated by a Gaussian type histogram. The Gaussian was centered on the actual age category it belonged to and the standard deviation (σ) of the Gaussian was varied to check the performance of the supervised i-vector. The Gaussian was sampled at each age category in the soft label vector, x, to get its appropriate weight for that particular age class, i.

$$L_i = e^{\frac{-(x-i)^2}{2\sigma^2}} \quad (2)$$

The standard deviation, σ, captures the range of ages which sound similar. We have incorporated the notion that a speech from a specific age sounds similar to the neighboring age classes. Using different σ gives us control over the size of these neighbouring age groups.

Psychoacoustic studies [3] show that younger and elderly age groups are more distinctive than the mid-range age groups. This motivated us to assign a Gaussian distribution to σ such that σ is lower for younger and elder age classes and higher for the mid-range age classes. This is a challenging task due to limited pre-knowledge and extensive parameter tuning. This in turn motivates us to construct and evaluate data driven models.

3.3. Data Driven Labels

The motive behind constructing data-driven labels is to discern how the i-vector captures the variability of age. Once we know how different one age is with respect to all the other ages, then we could embed this information in the i-vector training by using supervised i-vector concept to further increase the performance. First we explore methods to analyze the variability of age modeled in the i-vector space and then further use this information in our label matrix L to achieve better performance.

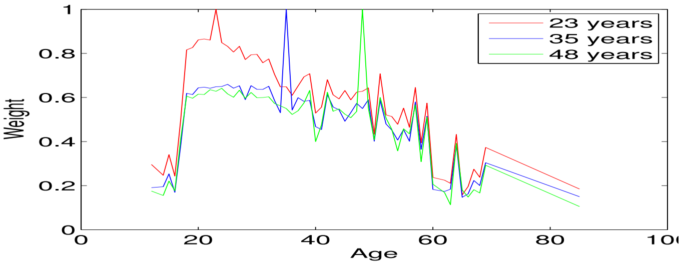

Fig. 3. *Data Driven Labels for 3 age groups*

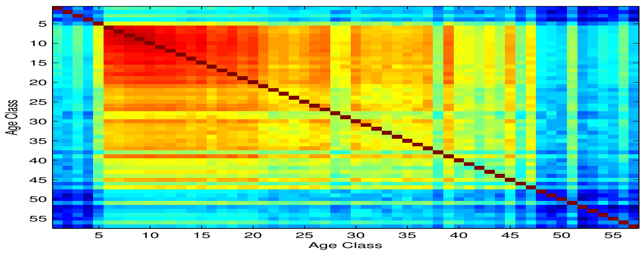

Fig. 4. *Confusion Matrix for the 2-pass data driven system*

To achieve this, we compute the mean i-vector for each speaker. This gives us a single i-vector representation for each speaker. Then the mean i-vector for each age class is computed on top of those speaker i-vectors. This is the i-vector representation for that age class. This is done for all age classes present in the training dataset. Finally, the euclidean distance between the features of each age class and all the other classes are computed. This gives a confusion matrix (with zeroes as the diagonal elements) of dimension M x M, where M is the number of unique age classes in the training dataset. The matrix is then normalized and each element in the matrix subtracted from 1. Hence we obtain a matrix having ones as the diagonal elements indicating highest weight assigned to the age class to which it belongs.

3.3.1. 2-pass data driven system

We run 2-passes for training, once using the traditional baseline i-vector, which is used to calculate the datadriven labels and then finally use this label for supervised training during the 2nd pass. Fig. 4 depicts the confusion matrix for our training data. This information can be viewed as the perceived age for each age class by the machine. We replace the chronological labels with the perceived ones as the label matrix in the supervised i-vector training. Fig. 3 shows the perceived age by machine for 3 different age classes (23, 35 and 48 years) for our training data. One can see that the age from young adults (23) are more confused with their neighbors.

3.3.2. 1-pass data driven system

To reduce the complexity and time required for 2-passes, we use adaptive datadriven labeling. The label matrix L adapts during each iteration of Expectation Maximization. Initially for the 1st iteration, since we don't have an estimate for the label matrix, we need to emulate the traditional i-vector using the supervised version. This can be achieved by initializing the covariance of the label reconstruction to infinity. Later, after the expectation step of the algorithm we calculate the datadriven labels for that iteration which will be used as the label vector L during the next iteration.

4. EXPERIMENTS

4.1. Database

The NIST SRE 2010 data [14] was used for training and the NIST SRE 2008 data was adopted for testing the SVR based estimation. The experiment was carried only for male speakers of the NIST 2010 and 2008 databases. The NIST 2010 set consists of 2611 utterances of male speakers with 49 unique age classes ranging from 19 to 87 years and NIST 2008 consists of 2147 utterances of male speakers with 52 unique age classes ranging from 17 to 80 years. The total variability matrix construction used 12060 utterances from male speakers (NIST 2010 (2611) + Switchboard Data (9449)). It is worth noting that the SVR training and testing data are the same as [6]. But we did not add NIST 04 05 06 data into the i-vector or supervised i-vector training as [6] did due to the absence of the speaker age labels.

4.2. Experimental Setup

For MFCC feature extraction, a 25ms Hamming window with 10ms shifts was adopted. A 36 dimensional MFCC feature vector was extracted consisting of 18 MFCC coefficients and their first derivatives. We used gender-dependent UBMs containing 1024 Gaussians and trained from NIST 2004 and NIST 2005 data. For our experiments we set the dimension of the i-vector to 400 to compare results with [6]. For regression, libsvm [15] was used for nu-SVR with Radial basis function as the kernel.

4.3. Performance Evaluation

In [6] Mean Absolute Error (MAE) is used to evaluate the performance of the system. The MAE is calculated as follows:

$$MAE = \frac{1}{Q}\sum_{q=1}^{Q}|\hat{y}_q - y_q| \tag{3}$$

where $\hat{y}_q$ and y_q are the predicted and the actual age from the q^{th} utterance of the testing data, respectively. Q is the total number of utterances in the testing data. Along with MAE, we also use Pearson's linear correlation coefficient (CORR) as an evaluation parameter for our experiments. CORR is a more reliable parameter compared to MAE for regression problems [16].

5. RESULTS AND DISCUSSIONS

Table 1 shows the result for the simplified versions of the i-vector and the supervised i-vector framework. Table 2 shows the result for the full versions of the i-vector and the supervised i-vector framework. For each version, we compare the performance of supervised i-vectors with our proposed label designs with respect to the baseline traditional i-vector in each of these 2 cases.

5.1. Simplified i-vector

Simplified version decreases the complexity and takes about 0.78% time of the full version with minimal loss in performance. Using the simplified i-vector reduces the performance by a relative 1.7%

Table 1. *Performance of simplified versions of i-vectors.*

Method	Label Design	LDA	WCCN	MAE	CORR
SIM IV	-	×	×	8.4671	0.4545
SIM IV	-	✓	×	8.4919	0.4605
SIM IV	-	×	✓	8.3797	0.4516
SIM SUP	B	×	×	8.5016	0.4476
SIM SUP	B	✓	×	9.0616	0.4388
SIM SUP	B	×	✓	**8.1962**	**0.4742**
SIM SUP	G1	×	×	8.2899	0.4838
SIM SUP	G1	✓	×	8.4387	0.4662
SIM SUP	G1	×	✓	8.381	0.4834
SIM SUP	G2	×	×	8.3149	0.4763
SIM SUP	G2	✓	×	8.4368	0.465
SIM SUP	G2	×	✓	8.2284	0.4826
SIM SUP	2P-DD	×	×	8.3785	0.4625
SIM SUP	2P-DD	✓	×	8.6059	0.4490
SIM SUP	2P-DD	×	✓	8.2805	0.4616
SIM SUP	1P-DD	×	×	8.3019	0.4711
SIM SUP	1P-DD	✓	×	8.2855	0.4786
SIM SUP	1P-DD	×	✓	**8.2061**	**0.4792**

SIM IV: Simplified i-vector, SIM SUP: Simplified Supervised i-vector 1P-DD: 1-pass Datadriven, 2P-DD: 2-pass Datadriven, G1: Gaussian σ=3, G2: Gaussian σ=N(30,10), B: Binary

Table 2. *Performance of Full versions of i-vectors.*

Method	Label Design	LDA	WCCN	MAE	CORR
IV	-	×	×	8.3253	0.4658
IV	-	✓	×	8.3176	0.4653
IV	-	×	✓	8.299	0.4601
SUP	B	×	×	8.4654	0.4689
SUP	B	✓	×	8.5556	0.4426
SUP	B	×	✓	**8.1257**	**0.4813**
SUP	G1	×	×	8.3016	0.472
SUP	G1	✓	×	8.5637	0.446
SUP	G1	×	✓	8.2713	0.4818
SUP	2P-DD	×	×	8.6242	0.4476
SUP	2P-DD	✓	×	8.6973	0.4344
SUP	2P-DD	×	✓	8.2675	0.4648
SUP	1P-DD	×	×	8.2618	0.4675
SUP	1P-DD	✓	×	8.3226	0.4643
SUP	1P-DD	×	✓	**8.1879**	**0.4776**

IV: i-vector, SUP: Supervised i-vector

(MAE) and 2.42% (CORR) as compared to the traditional i-vector. The traditional i-vector and the simplified i-vector results are used as the baseline for evaluating the supervised i-vector and the simplified supervised i-vector respectively.

5.2. Simplified Supervised and Supervised i-vectors

5.2.1. Binary label modeling

For both the supervised i-vector and the simplified supervised i-vector, the binary label design did not improve the performance without WCCN. This might be because our focused age estimation task is a regression task and hence directly assigning binary labels in the i-vector training may not lead to the improvement in regression. However, it makes the supervised i-vector more suitable for WCCN which is also based on discrete age classes. We can observe that applying the backend WCCN improved the MAE by 3.2% and CORR by 4.33% as compared to the baseline simplified i-vector and 2.4% (MAE) and 3.33% (CORR) as compared to the full version of the baseline i-vector.

5.2.2. Gaussian label modeling

For both the supervised i-vector and the simplified supervised i-vector, the Gaussian label designs show significant increase in performance as compared to the binary labels as well as the baseline systems before WCCN which indicates the effectiveness of adding age information in the i-vector training. Different values of σ are used to evaluate the Gaussian label design. Best performance was obtained using $\sigma = 3$ (see Table 1.). An increased performance of 0.28% (MAE) and 1.33% (CORR) is obtained for the supervised i-vector and 2.1% (MAE) and 6.44% (CORR) is obtained for the simplified supervised i-vector. Further improvement of 0.65% (MAE) and 3.43% (CORR) is achieved for the supervised i-vector over the baseline i-vector when combining WCCN.

5.2.3. Data Driven Labels

The data driven labels outperform both the binary and the Gaussian label design and has the best results as compared to the full-version baseline systems without intersession compensation. The 1-pass system outperforms the 2-pass datadriven system in both the simplified and the full versions. This is because in the 1-pass system we are incorporating the supervised system to calculate the confusion matrix, whereas in the 2-pass system the confusion matrix is calculated using traditional i-vectors.

For the 1-pass the best results are obtained using WCCN with an increase in 1.65% (MAE) and 2.53%(CORR). For the simplified version an increase in performance 3.08% (MAE) and 5.43% (CORR) is observed.

For the 2-pass the best results are obtained using WCCN with an increase in 0.69% (MAE). For the simplified version an increase in performance 2.2% (MAE) and 1.56% (CORR) is observed.

5.3. Intersession Compensation

Variability compensation techniques such as LDA and WCCN were applied. In all cases LDA did not help for both MAE and CORR (see Table 1, 2) whereas WCCN improves the performance in all cases of the supervised i-vectors. The performance of WCCN is highest in Binary labels. This is because the Binary label design vector performs hard discrimination between age classes whereas the Gaussian label/Data Driven design vector assigns soft labels between age classes. Thus, there is no clear individual class assigned which leads to relatively poor performance of WCCN for the Gaussian/Data Driven label designs with respect to Binary Labels.

6. CONCLUSIONS

In this paper, we showed that the supervised i-vector outperforms the i-vector baseline for the speaker age regression problem by a statistically significant margin ($p < 0.05$) when using MAE and CORR as an evaluation parameter. Novel label vector modeling techniques have been proposed to improve the performance of the supervised i-vector modeling and better reflect the gradient nature of age estimation. The performance of the proposed system is compared with the current state of the art method with promising results.

The 1 pass data driven labels (supervised i-vector framework) gave the best results for systems without inter-session compensation. Using WCCN for intersession compensation, a further increase in performance was obtained for all the label designs in the supervised and simplified supervised i-vector framework. The best performance using WCCN was obtained using the Binary Label design improving the MAE by 3.2% and CORR by 4.33% as compared to the baseline simplified i-vector and 2.4% (MAE) and 3.33% (CORR) as compared to the full version of the baseline i-vector.

7. REFERENCES

[1] S. Narayanan and P.G. Georgiou, “Behavioral signal processing: Deriving human behavioral informatics from speech and language,” in *Proceedings of the IEEE*, 2013, vol. 101, pp. 1203–1233.

[2] D. Elenius and M. Blomberg, “Comparing speech recognition for adults and children,” in *in Proceedings of FONETIK*, 2004.

[3] L. Cerrato, M. Falcone, and A. Paoloni, “Subjective age estimation of telephonic voices,” in *Speech Communication*, 2000, vol. 31, pp. 107 – 112.

[4] M. Li, K. J. Han, and S. Narayanan, “Automatic speaker age and gender recognition using acoustic and prosodic level information fusion,” in *Comput. Speech Lang.*, Jan. 2013, vol. 27, pp. 151–167.

[5] B.R. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, “Paralinguistics in speech and language–state-of-the-art and the challenge,” 2012.

[6] Bahari M.H., McLaren M., Van hamme H., and Van Leeuwen D., “Age estimation from telephone speech using i-vectors,” in *Annual conference of the International Speech Communication Association (ISCA)*. Interspeech, 2012, vol. 13, pp. 506–509.

[7] F. Lingenfelser, J. Wagner, T. Vogt, J. Kim, and E. André, “Age and Gender Classification from Speech Using Decision Level Fusion and Ensemble Based Techniques,” in *INTERSPEECH*, 2010, pp. 2798–2801.

[8] N. Dehak, P.A.T. Carrasquillo, D. Reynolds, and R. Dehak, “Language recognition via ivectors and dimensionality reduction,” in *Annual conference of the International Speech Communication Association (ISCA)*. Interspeech, 2011.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” in *Audio, Speech, and Language Processing, IEEE Transactions on*, 2011, vol. 19, pp. 788–798.

[10] T. Bocklet, A. Maier, J.G. Bauer, F. Burkhardt, and E. Noth, “Age and gender recognition for telephone applications based on gmm supervectors and support vector machines,” in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 1605–1608.

[11] M. Li, A Tsiartas, M.V. Segbroeck, and S. Narayanan, “Speaker verification using simplified and supervised i-vector modeling,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7199–7203.

[12] R.O. Duda, P.E. Hart, and D.G. Stork, “Pattern classification (2nd edition),” 2000, Wiley-Interscience.

[13] A.O. Hatch, S.S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for svm-based speaker recognition,” in *Proc. of ICSLP*, 2006, p. 14711474.

[14] National Institute of Standards and Technology, “The nist year 2010 speaker recognition evaluation,” Data available at http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html.

[15] C.C. Chang and C.J. Lin, “LIBSVM: A library for support vector machines,” in *ACM Transactions on Intelligent Systems and Technology*, 2011, vol. 2, pp. 27:1–27:27, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[16] R.V. Hogg and E.A. Tanis, “Probability and statistical inference,” 2004.