Multi-way Regression for Age Prediction Exploiting Speech and Face Image Information

Evangelia Pantraki, and Constantine Kotropoulos
Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki 54124, GREECE
Email: epantrak@csd.auth.gr, costas@aiia.csd.auth.gr

***Abstract*—In this paper, the problem of age estimation is addressed based on two modalities: speech utterances and speakers' face images. The proposed age estimation framework employs the Shifted Covariates REgression Analysis for Multi-way data (SCREAM) model, which combines Parallel Factor Analysis 2 and Principal Covariates Regression. SCREAM is able to extract a few latent variables from multi-way data and compute regression coefficients. Initially, biologically inspired features are extracted from speech utterances and face images and are suitable feature matrices are created to be fed to the multi-way SCREAM model. For bimodal age estimation, the visual and aural features are appropriately combined in a single matrix for each person. Experimental results demonstrate the profit of combining the two modalities. The performance admitted by the multi-way regression for age estimation is also measured on the benchmark face image dataset FG-NET. The proposed method is found to be competitive to state-of-the-art age estimation methods.**

I. Introduction

Automatic human age estimation is a very challenging task that has attracted great research interest over the years. Age estimation is useful in fields such as biometrics, security control and surveillance monitoring, forensics, and electronic customer relationship management [1]. In this paper, bimodal age estimation is proposed that is based on both face images and speech utterances. When both speech and face images are available, the consideration of both modalities could potentially improve age estimation accuracy, especially if one of the two modalities provides noisy input. For example, the combination of aural and visual cues could be useful in the analysis of video sequences of crime scenes captured by surveillance cameras/microphones. In such cases, either speech signals or face images of suspects may be corrupted with noise or occlusions, hence the combination of speech and face data could yield better age estimates of the persons appearing in the video, supporting in that way the process of suspect identification.

Most existing methods for automatic age estimation exploit face image information. The release of benchmark FG-NET Ageing Database [2] in 2004 supported significantly the research on facial age estimation. A comprehensive overview of research activities in facial age estimation can be found in [1]. Facial ageing patterns are extracted for automatic age estimation in [3], while a hierarchical approach is proposed in [4]. This approach initially performs age group classification and subsequently, Support Vector Regression (SVR) is applied to predict the final age. Biologically inspired features (BIFs) based on Gabor filters are deployed for age estimation in [5]. In [6], a framework for age estimation via face image analysis is proposed that includes face detection, discriminative manifold learning, and multiple linear regression.

Many research efforts have focused on speaker age estimation based solely on speech signal. In [7], Gaussian mixture model supervectors are employed as features, weighted-pairwise principal components analysis is applied for dimensionality reduction and SVR is employed for age regression. In [8], Linear Discriminant Analysis is performed to reduce the dimension of i-vectors and SVR is utilized for automatic age estimation. In [9], speech utterances are modeled using i-vectors and least squares SVR is applied to estimate the age of speakers.

This paper extends the work on age interval prediction based on speech utterances reported in [10]. In [10], Parallel Factor Analysis 2 (PARAFAC2) [11] was applied to an irregular third-order tensor for age group classification. Here, a well known multi-way regression method, namely the Shifted Covariates REgression Analysis for Multi-way data (SCREAM) [12], is employed for age estimation. SCREAM is based on a combination of PARAFAC2 and Principal Covariates Regression (PCovR). Experiments are conducted for age estimation based solely on speech, solely on face images, and on both speech utterances and face images. To this end, the Trinity College Dublin Speaker Ageing database (TCDSA) [13] is supplemented with face images of the speakers, which are contemporary to their speech recordings. Moreover, experiments are conducted on the benchmark FG-NET Aging dataset in order to illustrate better the performance of the proposed age estimation framework. Experimental results evidently demonstrate the proposed framework competency in age estimation.

II. Datasets

The proposed age estimation framework is initially applied to the longitudinal TCDSA database [13] for speech-based age prediction. The database contains recordings spanning a year range per speaker varying between 30 and 60 years at irregular intervals between 1 to 10 years. The duration of speech recordings in the TCDSA database varies from 25 seconds to 35 minutes. The database includes a different number of recordings per speaker, varying from 4 to 47 recordings per

speaker. The total number of speakers included in the TCDSA dataset is 26, including 15 males and 11 females.

In order to perform bimodal age estimation, face images were collected for each speaker of the dataset by locating publicly available visual material portraying him/her. Effort has been devoted so that the face images were captured close to the speakers' age. Since the exact matching is difficult, a 3-year tolerance is allowed between the age of a person when his/her face was captured and the age associated to his/her utterance. Such a 3-year tolerance is not expected to affect the exactness of speech and face image matching due to the gradual progression of ageing process.

A total duration of 30 seconds is kept from each recording or less if the recording's duration is shorter than 30 seconds. If many face images of the person at the age of the speech recording have been collected, more than one segments of 30 seconds long are kept. A total of 227 recordings could be matched with contemporary face images. Finally, the total number of speakers included in the extended TCDSA audio-visual dataset was 25, including 14 males and 11 females.

The collected face images were resized to 60×60 pixels and the face was cropped in order to remove background. All face images were converted to grayscale. Some examples of the collected face images for four speakers of the extended TCDSA dataset are depicted in Figure 1. Pose and illumination vary greatly over the collected face images, as can be observed by the sample face images depicted in Figure 1. To the best of our knowledge, the extended TCDSA dataset that combines age separated speech samples and face images, is a unique dataset that supports bimodal age estimation experiments using aural and visual features.

Fig. 1. Face images depicting four speakers of the extended TCDSA dataset at ascending ages.

A second set of experiments was conducted on the FG-NET benchmark dataset, which comprises of 1002 face images that belong to 82 unique persons (48 male and 34 female) at various ages [2]. Similar to the TCDSA database, face images were resized to 60×60 pixels and the face was cropped in order to remove background.

III. Proposed Method

The proposed framework treats age estimation problem as a regression problem. Here, our goal is to exploit the multi-way regression method SCREAM [12] for age estimation. The first step is to perform feature extraction from speech utterances and face images in order to obtain discriminative feature representations. The SCREAM regression model requires multi-way data, therefore the features extracted from face images and speech utterances need to be re-arranged in matricized form. The second step includes the training of a SCREAM regression model on speech and face image feature matrices in order to estimate age. For bimodal age estimation, the speech and face image feature matrices are appropriately combined in a single matrix, which is subsequently fed to a SCREAM regression model.

A. Feature extraction

Auditory cortical representations are extracted from speech utterances. These descriptors are inspired by the way sound is perceived and processed by the human auditory system [14]. Auditory cortical representations are actually a four-dimensional (4D) representation of time, frequency, rate, and scale. The auditory cortical representations extracted for each frame are averaged across time and the resulting 3D representations for each speech recording are obtained (frequency channels $\times$ rates $\times$ scales). By appropriately unfolding the 3D representation across frequency dimension, a matrix $\mathbf{X_1} \in \mathbb{R}_+^{(rates \times scales) \times frequency\_channels}$ is obtained. Following [15], 10 rates, 6 scales and 128 filters are employed, which cover 8 octaves between 44.9 Hz and 11 kHz. Therefore, each speech recording is represented by a matrix $\mathbf{X_1} \in \mathbb{R}_+^{60 \times 128}$.

BIFs are extracted from each face image following the procedure for human age estimation proposed in [5]. These features are actually a pyramid of Gabor filters and are similar to the way the human visual system processes visual stimuli. In total, 8 bands and 12 orientations were chosen for the applied Gabor filters. For each scale band, the pooling grids are 6×6, 8×8, 10×10, 12×12, 14×14, 16×16, 18×18, and 20×20, respectively. Moreover, the allowed overlaps for each pooling grid are 3, 4, 5, 6, 7, 8, 9, and 10, respectively. Each scale band has a pair of adjacent filter sizes, therefore two maps are obtained for each orientation: one for each filter. The maximum operation "MAX" is used as a pooling filter on the two maps and the maximum map is obtained. In addition to the "MAX" pooling, a nonlinear standard deviation operation ("STD" operation) is proposed in [5] in order to capture the local variations of ageing. The "STD" operation is applied on the maximum map using the band's pooling grid. From the aforementioned procedure, a total of 1099 values are obtained for each orientation across bands for each 60×60 pixel image. More specifically, 20×20 values are obtained for the first band which has a 6×6 pooling grid and an overlap of 3, 15×15 values are obtained for the second band which has an 8×8 pooling grid and an overlap of 4, and so on. The 1099 feature values obtained for each orientation are arranged as the matrix of dimension 20×87 depicted in Figure 2. The remaining 641 entries in each matrix that do not correspond to feature values are filled with zeros. In total, 12 such matrices are created, one for each orientation.

Subsequently, the matrices are concatenated horizontally to yield a single matrix $\mathbf{X}_2 \in \mathbb{R}_+^{20\times 1044}$ for each face image.

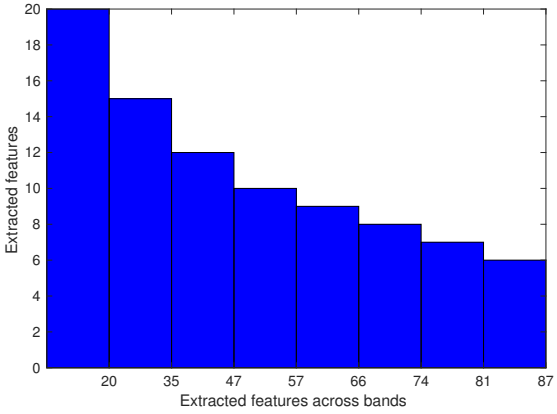

Fig. 2. BIFs across the 8 bands for one orientation.

In order to perform bimodal age estimation, the speech and face image feature matrices are suitably merged in a single matrix, which is subsequently fed to SCREAM regression model. More specifically, the speech feature matrix and the face image feature matrix are concatenated in a block diagonal matrix $\mathbf{X_3} \in \mathbb{R}_+^{80\times 1172}$.

B. SCREAM-based multi-way regression

SCREAM is a multi-way regression method proposed in [12]. The method is a combination of PARAFAC2 decomposition and PCovR. SCREAM employs PARAFAC2 to decompose a multi-way data matrix $\mathcal{X}$ and PCovR to compute regression coefficients by minimizing a unified least squares criterion. PARAFAC2 is a multi-way generalization of the Singular Value Decomposition (SVD), which can be applied to a collection of matrices having the same number of columns, but different number of rows [16]. PARAFAC2 assumes a common latent structure in the input tensor $\mathcal{X}$ and performs dimensionality reduction by extracting latent variables. By combining PARAFAC2 with PCovR, a regression model, coined as SCREAM, has been developed that is able to provide latent variables useful for predictive models.

The proposed framework exploits the multi-way regression model SCREAM for age estimation. A SCREAM model is trained on a three-way array $\mathcal{X} \in \mathbb{R}^{I\times J\times K}$, where K is the number of samples. The kth sample is denoted by $\mathbf{X}_k \in \mathbb{R}^{I\times J}$. For the proposed age estimation framework, each sample matrix $\mathbf{X}_k$ corresponds to the matricized feature representations of speech recordings and face images. More specifically, for age estimation based solely on speech, the speech feature matrices $\mathbf{X}_1 \in \mathbb{R}_+^{60\times 128}$ are combined in a three-way array $\mathcal{X}_1 \in \mathbb{R}_+^{60\times 128\times K}$, where K is the number of speech recordings. For age estimation based solely on face images, the face image feature matrices are combined in a three-way array $\mathcal{X}_2 \in \mathbb{R}_+^{20\times 1044\times K}$. Finally, for bimodal age estimation, the three-way array $\mathcal{X}_3 \in \mathbb{R}_+^{80\times 1172\times K}$ that combines speech and face image feature representations is utilized.

The SCREAM model combines the loss function of PARAFAC2 decomposition with the loss function of a regression model into a single optimization problem. A scalar parameter α, $0 \leq \alpha \leq 1$, is utilized to compensate between the goodness of fit of the decomposition of $\mathcal{X}$ and the error of the regression of $\mathbf{y} \in \mathbb{R}_+^{K\times 1}$. The vector $\mathbf{y}$ comprises the person ages. During the training phase, the ground truth person ages are the elements of $\mathbf{y}$. During the test phase, $\mathbf{y}$ contains the predicted ages returned by the SCREAM model. That is, $\mathbf{y}$ is the dependent regression variable of the SCREAM model. The SCREAM model seeks a solution to the optimization problem:

$$\underset{\mathbf{C},\,\mathbf{P},\,\mathbf{r}}{\operatorname{argmin}}\ \alpha\|\mathbf{X}-\mathbf{C}\mathbf{P}^{\intercal}\|_F^2+(1-\alpha)\|\mathbf{y}-\mathbf{C}\mathbf{r}\|_2^2. \qquad (1)$$

The first term of loss function (1) is actually the loss function of PARAFAC2:

$$\sum_{k=1}^{K}(\mathbf{X}_k-\mathbf{A}\mathbf{D}_k\mathbf{B}_k^{\intercal})^2=\|\mathbf{X}-\mathbf{C}\mathbf{P}^{\intercal}\|_F^2 \qquad (2)$$

where $\mathbf{X} \in \mathbb{R}_+^{K\times IJ}$ holds the appropriately unfolded three-way array $\mathcal{X} \in \mathbb{R}_+^{I\times J\times K}$ and $\mathbf{P}$ is a matrix holding the matrices $\mathbf{A}$ and $\mathbf{B}_k$ appropriately arranged. $\mathbf{A} \in \mathbb{R}^{I\times F}$ is the first mode loadings of an F-component PARAFAC2 model, where F is the number of latent variables extracted. The diagonal matrix $\mathbf{D}_k \in \mathbb{R}_+^{F\times F}$ holds the kth row of the third mode loadings $\mathbf{C} \in \mathbb{R}^{K\times F}$. The third mode is the sample mode. The matrix $\mathbf{B}_k \in \mathbb{R}^{J\times F}$ is the loadings for second mode for sample k. To achieve uniqueness, the square matrix $\mathbf{B}_k^{\intercal}\mathbf{B}_k=\mathbf{H}$ is kept constant over k [11]. The second term of loss function (1) is actually the regression error:

$$\|\mathbf{y}-\mathbf{C}\mathbf{r}\|_2^2 \qquad (3)$$

where $\mathbf{r} \in \mathbb{R}^{F\times 1}$ is a vector of regression coefficients. The matrix $\mathbf{C} \in \mathbb{R}^{K\times F}$ is used concurrently for fitting the PARAFAC2 model of $\mathcal{X}$ and for the regression problem of predicting $\mathbf{y}$. Therefore, it integrates the two minimization objectives. In order to ensure that the components of $\mathbf{C}$ are relevant for prediction, it is expressed as $\mathbf{C}=\mathbf{X}\mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{IJ\times F}$ is a weight matrix that ensures that $\mathbf{C}$ is in the row-space of $\mathbf{X}$.

Here, the SCREAM model is utilized for age estimation. By solving the optimization problem (1), a reduced dimension representation model of the three-way feature array $\mathcal{X}$ is obtained and concurrently a regression for age prediction is performed. SCREAM is a powerful regression technique and to our knowledge, it is the first time that it is exploited for age estimation.

IV. Experimental Evaluation

In order to assess the performance of the proposed framework in age estimation, experiments were conducted on the audio-visual TCDSA and the facial FG-NET datasets, described in detail in Section II. During the evaluation, the

Leave-One-Person-Out (LOPO) evaluation protocol was applied. Successively, the observations (speech recordings and face images) of each speaker were included into the test set, while the observations belonging to the remaining speakers of the dataset were used for training. LOPO defines $M = 25$ folds in the audio-visual TCDSA dataset, one for each person. Similarly, $M = 82$ folds were defined in the FG-NET dataset.

Since the audio-visual TCDSA dataset consists of 227 recordings and each speech feature matrix $\mathbf{X}_1$ has dimension 60×128, the 227 samples formulate the three-way speech feature matrix $\mathcal{X}_1 \in \mathbb{R}^{60\times128\times227}$. The three-way face image feature matrix $\mathcal{X}_2 \in \mathbb{R}^{20\times1044\times227}$ and the combined feature matrix $\mathcal{X}_3 \in \mathbb{R}^{80\times1172\times227}$ are formulated. For the FG-NET dataset, only the face image modality is available, therefore the SCREAM-based age estimation framework is only applied to the three-way face image feature matrix $\mathcal{X}_2 \in \mathbb{R}^{20\times1044\times1002}$, where $K = 1002$ the number of face images.

The Mean Absolute Error (MAE) was employed as a regression metric to assess the estimations made by the proposed method. MAE is the average of the absolute errors between the predicted age value and the actual age value. A range of different values for the number of latent variables F extracted by SCREAM and the parameter α were tested in order to find the most suitable values for each dataset.

The best MAE for the proposed age estimation framework in several experiments on the TCDSA dataset is summarized in Table I. Here, the modality exploited in each experiment is denoted as either speech, image or speech+image. For the speech modality, only aural features were utilized and the speech feature matrices were fed to the SCREAM model. Similarly, for image and speech+image modalities, only face feature matrices and combined speech and face feature matrices are utilized, respectively. The parameter values for the results presented in Table I are $\alpha = 0.6$ and $F = 1$ for the speech modality, $\alpha = 0.9$ and $F = 2$ for the image modality, and $\alpha = 0.8$ and $F = 1$ for the speech+image modality. It is seen from Table I that the best performance on age prediction was achieved based on the combination of speech and face image modalities. In addition, the proposed age estimation framework yielded a better MAE when it was based exclusively on speech features rather than when it was based exclusively on face image features.

TABLE I
MAE (YEARS) AT DIFFERENT AGE GROUPS ON TCDSA DATASET USING DIFFERENT MODALITIES.

Age estimation results - TCDSA				
Range	**#records**	**Speech**	**Image**	**Speech+Image**
21-29	31	24.03	25.63	15.65
30-39	28	15.64	16.22	11.28
40-49	38	7.52	7.03	8.12
50-59	47	4.87	6.95	8.96
60-69	40	10.45	10.57	11.19
70-79	29	18.48	17.15	16.44
80-89	14	28.48	24.76	31.33
Total	227	13.44	13.70	12.75

In order to examine whether the proposed age estimation framework is robust to noise corruption, we conducted the same experiments presented in Table I having added car idle noise to the speech recordings [17]. The experimental findings after the addition of noise to the speech recordings are presented in Table II. As expected, the MAEs for speech-based and bimodal age estimation have been increased, but bimodal age estimation still outperforms speech-based age estimation. Therefore, the inclusion of the face image modality increases the robustness of the bimodal age estimation framework. Interestingly, bimodal age estimation outperforms image-based age estimation, despite the addition of noise to the speech recordings. The ability of the proposed bimodal age estimation framework to deal effectively with noisy input of one modality is noteworthy, since in most real life applications involving audio and visual input, one of the two modalities is likely to be corrupted with noise.

TABLE II
MAE (YEARS) AT DIFFERENT AGE GROUPS ON TCDSA DATASET USING DIFFERENT MODALITIES AFTER THE ADDITION OF NOISE TO SPEECH RECORDINGS.

Age estimation results - TCDSA				
Range	**#records**	**Speech**	**Image**	**Speech+Image**
21-29	31	24.87	25.63	24.76
30-39	28	15.75	16.22	15.41
40-49	38	6.61	7.03	6.03
50-59	47	4.38	6.95	3.86
60-69	40	11.5	10.57	11.16
70-79	29	19.57	17.15	18.66
80-89	14	27.85	24.76	27.5
Total	227	13.60	13.70	13.13

The best MAE for the proposed age estimation framework in several experiments on the FG-NET dataset is summarized in Table III. Here, only the face modality is available, therefore only face image-based age estimation is performed. The parameter values of the proposed method for the results presented in Table III are $\alpha = 0.4$ and $F = 3$. From Table III, it is seen that the proposed method performance on age estimation is promising, when compared to other age estimation techniques applied to the benchmark FG-NET dataset.

TABLE III
MAE (YEARS) AT DIFFERENT AGE GROUPS ON FG-NET DATASET.

Age estimation results - FG-NET						
Range	**#records**	**Proposed**	**BIF [5]**	**RUN [18]**	**QM [19]**	**MLP [19]**
0-9	371	6.78	2.99	2.51	6.26	11.63
10-19	339	5.40	3.39	3.76	5.85	3.33
20-29	144	6.73	4.30	6.38	7.10	8.81
30-39	79	10.52	8.24	12.51	11.56	18.46
40-49	46	17.69	14.98	20.09	14.80	27.98
50-59	15	28.04	20.49	28.07	24.27	49.13
60-69	8	38.09	31.62	42.50	37.38	49.13
Total	227	7.67	4.77	5.78	7.57	10.39

V. Conclusions and future work

In this paper, we have addressed age estimation by employing the multi-way regression method SCREAM. SCREAM is utilized for age prediction based on speech recordings and face images. Biologically inspired features are extracted from speech utterances and face images and are suitably arranged to feature matrices so that they can be fed to the multi-way SCREAM model. Moreover, the aural and visual feature matrices are combined in a single matrix, enabling bimodal age estimation using SCREAM. Experimental results on the audio-visual TCDSA dataset demonstrate that the proposed age estimation framework achieved its best performance when both modalities were utilized. The age estimation based solely on speech is slightly more accurate than that based solely on face images. Experiments on the benchmark FG-NET dataset demonstrate that the proposed age estimation framework performance is competitive to other age estimation techniques. The proposed method possesses a unique advantage than the state-of-the-art techniques: it depends of **few** parameters (i.e., 2-3 latent variables and another 2-3 regression coefficients), which makes it a first candidate for big data applications if properly implemented. Future work will focus on expanding the SCREAM model so that it allows the integration of information from two modalities directly into (1) and solving the resulting optimization problem.

Acknowledgments. The authors are grateful to Dr. F. Kelly and Prof. R. Bro for having shared with them the TCDSA dataset and the code of SCREAM, respectively.

References

[1] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1955–1976, November 2010.

[2] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "Overview of research on facial ageing using the FG-NET ageing database," *IET Biometrics*, vol. 5, no. 2, pp. 37–46, June 2016.

[3] X. Geng, Z. H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2234–2240, December 2007.

[4] H. Han, C. Otto, and A. K. Jain, "Age estimation from face images: Human vs. machine performance," in *Proc. IEEE Int. Conf. Biometrics*, Madrid, Spain, June 2013, pp. 1–8.

[5] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, Miami, Florida, US, June 2009, pp. 112–119.

[6] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, June 2008.

[7] G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1975–1985, September 2011.

[8] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Shanghai, China, March 2016, pp. 5040–5044.

[9] M. H. Bahari, M. McLaren, H. V. Hamme, and D. A. V. Leeuwen, "Speaker age estimation using i-vectors," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 99–108, September 2014.

[10] E. Pantraki, C. Kotropoulos, and A. Lanitis, "Age interval and gender prediction using PARAFAC2 applied to speech utterances," in *Proc. 4th Int. Workshop on Biometrics and Forensics*, Limassol, Cyprus, March 2016, pp. 1–6.

[11] R. A. Harshman, "PARAFAC2: Mathematical and technical notes," *UCLA Working Papers in Phonetics*, vol. 22, pp. 30–47, 1972.

[12] F. Marini and R. Bro, "SCREAM: A novel method for multi-way regression problems with shifts and shape changes in one mode," *Chemometrics and Intelligent Laboratory Systems*, vol. 129, pp. 64–75, November 2013.

[13] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification with long-term ageing data," in *Proc. 5th IAPR Int. Conf. Biometrics*, New Delhi, India, March 2012, pp. 478–483, http://www.mee.tcd.ie/~sigmedia/Research/SpeakerVerification.

[14] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, May 2006.

[15] Y. Panagakis, C. L. Kotropoulos, and G. R. Arce, "Music genre classification via joint sparse low-rank representation of audio features," *IEEE/ACM Trans. Audio, Speech, and Language Processings*, vol. 22, no. 12, pp. 1905–1917, December 2014.

[16] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, August 2009.

[17] P. C. Loizou, *Speech enhancement: Theory and practice (2nd edn.)*. CRC Press, 2013.

[18] S. Yan, H. Wang, X. Tang, and T. S. Huang, "Learning auto-structured regressor from uncertain nonnegative labels," in *Proc. IEEE Int. Conf. Computer Vision*, Rio de Janeiro, Brazil, October 2007, pp. 1–8.

[19] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 621–628, February 2004.