AGE INTERVAL AND GENDER PREDICTION USING PARAFAC2 APPLIED TO SPEECH UTTERANCES

Evangelia Pantraki[*], *Constantine Kotropoulos*[†]

Aristotle University of Thessaloniki
Thessaloniki 54124, GREECE
{pantraki@|costas@aiia}.csd.auth.gr

Andreas Lanitis[‡]

Cyprus University of Technology
3040 Limassol, Cyprus
andreas.lanitis@cut.ac.cy

ABSTRACT

Important problems in speech soft biometrics include the prediction of speaker's age or gender. Here, the aforementioned problems are addressed in the context of utterances collected during a long time period. A unified framework for age and gender prediction is proposed based on Parallel Factor Analysis 2 (PARAFAC2). PARAFAC2 is applied to a collection of three matrices, namely the speech utterance-feature matrix whose columns are the auditory cortical representations, the speaker age matrix whose columns are indicator vectors of suitable dimension, and the speaker gender matrix whose columns are proper indicator vectors associated to speaker's gender. PARAFAC2 is able to reduce the dimensionality of the auditory cortical representations by projecting these representations onto a semantic space dominated by the age and the gender concepts, yielding a sketch (i.e., a feature vector of reduced dimensions). To predict speaker's age interval associated to a test utterance, the speech utterance sketch is pre-multiplied by the left singular vectors of the speaker age matrix. To predict the gender of the speaker who uttered any test utterance, the speech utterance sketch is pre-multiplied by the left singular vectors of the speaker gender matrix. In both cases, a ranking vector is obtained that is exploited for decision making. Promising results are demonstrated, when the aforementioned framework is applied to the Trinity College Dublin Speaker Ageing Database.

Index Terms— Speaker biometrics, speaker ageing, PARAFAC2.

1. INTRODUCTION

The acoustic changes of the human voice as a result of ageing, known as *vocal ageing*, have been thoroughly studied in [1–3]. For example, the respiratory system is affected by the decreasing rate and strength of muscle contraction. The primary anatomic changes of larynx are the ossification of cartilages and the atrophy of muscle tissue. Furthermore, the loss of functionality of the tongue and facial muscles affect the supralaryngeal system [4].

Biometric templates are actually snapshots of biometric characteristics captured at a particular time instant [5]. In a verification scenario, the decision is influenced heavily whenever a time lapse between the enrolment and the verification exists [6]. Ageing becomes an important issue in recognition, as well [7]. Compensating for ageing in face verification has received significantly more attention than in speaker verification. Motivated by the release of publicly available databases for face verification, such as the FG-NET Ageing database [8] or the MORPH database [9], similar initiatives undertaken by the speech research community have led to the release of the Greybeard - Voice and Ageing Database distributed by the Linguistic Data Consortium [10], the University of Florida Vocal Aging Database [11], and the longitudinal Trinity College Dublin Speaker Ageing (TCDSA) [12].

An evaluation of speaker verification on the TCDSA database with a Gaussian Mixture Model - Universal Background Model (GMM-UBM) system revealed that the verification scores of genuine speakers decreased progressively as the time span between training and testing increased, while the imposter scores were less affected [12]. The addition of temporal information to the mel frequency cepstral coefficients (MFCCs) caused an increase in the rate of degradation [4]. However, at time-lapse of 30 years, vocal ageing caused significant problems in forensic automatic speaker recognition [13]. Combining ageing information with quality measures and scores from the GMM-UBM system, a decision boundary was created in the score-ageing-quality space [14]. By reducing the variability related to non-ageing, the accuracy of long-term ageing-dependent decision boundary improved. Eigenageing compensation was proposed to adapt a speaker model to a test sample based on a vocal ageing subspace [15]. The performance of the i-vector system in terms of both discrimination and calibration was found to degrade progressively as the absolute age difference between the training and test samples increased [16].

Here, we are interested in predicting the chronological

[*] Supported by a Short Term Scientific Mission grant from COST Action IC1106 "Integrating Biometrics and Forensics for the Digital Age."

[†] Supported by the COST Action IC 1106.

[‡] Supported also by the COST Action IC 1106.

age and the gender from utterance-level acoustic features. Three novel systems combining short-term cepstral features and long-term features for speaker age recognition were compared to each other in [17]. A system combining GMMs using frame-based MFCCs and Support-Vector-Machines using long-term pitch was found to perform best. A parallel phone recognizer was found to yield a comparable performance to human listeners in automatic age and gender classification using seven classes on a telephony speech task, while loosing performance on short utterances [18]. By adding prosodic, pitch, and formant features to the MFCCs, a relative reduction of the mean absolute error in speaker age estimation was reported in [19].

A novel framework for age and gender prediction is proposed that is based on Parallel Factor Analysis 2 (PARAFAC2) [20]. In the training phase, the starting point is to form an irregular third-order tensor (or more precisely hypermatrix) having three slices. The first slice is the speech utterance feature matrix, whose columns are the features extracted from speech utterances. Contrary to the majority of related methods, which resort to MFCCs, the auditory cortical representations are computed from each utterance. These features are based on spectrotemporal modulations [21] and their derivation is motivated by the human auditory system. The second slice is the speaker age matrix whose columns are indicator vectors of suitable dimension associated to speaker's age. The third slice is the speaker gender matrix, whose columns are indicator vectors of proper dimension associated to speaker's gender. The choice regarding the dimensions of the age and gender indicator vectors will be discussed later on. PARAFAC2 is applied to the aforementioned irregular third-order tensor so that the semantic similarities between the age and gender annotations of the utterances drive the extraction of meaningful feature vectors of reduced dimensions referred to as *sketches* hereafter. The reasoning behind this approach is that PARAFAC2 represents the feature vector and the associated age and gender vectors as linear combinations of basis vectors with coefficients taken from the same vector space. The left singular vectors of the speech utterance feature matrix span a lower dimensional semantic space dominated by the age and gender information. Any auditory cortical representation vector extracted from a test utterance is projected onto this semantic space first in order to obtain a test sketch. To predict speaker's age interval associated to a test utterance, the test sketch is pre-multiplied by the left singular vectors of the speaker age matrix. To predict the gender of the speaker who uttered any test recording, the test sketch is pre-multiplied by the left singular vectors of the speaker gender matrix. In both cases, a ranking vector is derived that is exploited for decision making. Promising results are demonstrated when the aforementioned framework is applied to the TCDSA Database, using a 2-fold cross validation protocol.

The paper is organized as follows. Section 2 begins with basic notation. The proposed joint age and gender prediction framework, which is based on PARAFAC2, is detailed next. Experimental results are demonstrated in Section 3, and conclusions are drawn in Section 4.

2. JOINT AGE ESTIMATION AND GENDER PREDICTION

2.1. Notation

Tensors are considered as the multidimensional equivalent of matrices (i.e., second-order tensors) and vectors (i.e., first-order tensors) [22]. Throughout the paper, tensors are denoted by boldface Euler script calligraphic letters (e.g. $\boldsymbol{\mathcal{X}}$), matrices are denoted by uppercase boldface letters (e.g., $\mathbf{U}$), vectors are denoted by lowercase boldface letters (e.g., $\mathbf{u}$), and scalars are denoted by lowercase letters (e.g., u). $\|.\|_F$ denotes the Frobenius matrix norm, while $\mathbf{B}^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{B}$. Let $\mathbb{Z}$ and $\mathbb{R}$ denote the set of integer and real numbers, respectively. A third-order real-valued tensor $\boldsymbol{\mathcal{X}}$ is defined over the tensor space $\mathbb{R}^{I_1 \times I_2 \times I_3}$, where $I_n \in \mathbb{Z}$ and $n = 1, 2, 3$. Each element of $\boldsymbol{\mathcal{X}}$ is addressed by 3 indices, i.e., $x_{i_1 i_2 i_3}$. Hereafter, the operations on tensors are expressed in matricized form [22].

2.2. Proposed method

PARAFAC is a multi-way generalization of the singular value decomposition (SVD) [23]. PARAFAC2 [20] is a variant of PARAFAC, which relaxes some of PARAFAC constraints. That is, while PARAFAC applies the same factors across a set of matrices, PARAFAC2 applies the same factor along one mode. The aforementioned relaxation allows the other factor matrices to vary, enabling the application of PARAFAC2 to a collection of matrices having the same number of columns, but different number of rows [22]. Such a collection forms the slices of an irregular third-order tensor. Another important characteristic of PARAFAC2 is its ability to overcome the weakness of conventional supervised subspace learning algorithms to handle multi-labelled data. Due to these characteristics, PARAFAC2 has emerged as an appealing method for multi-label classification. It has been applied successfully to feature extraction and multi-label classification of documents [24] and music tagging [25]. Here, our goal is to exploit the good decomposition properties of the PARAFAC2 to jointly predict speaker's age and gender.

A PARAFAC2 model is trained on an irregular third-order tensor $\boldsymbol{\mathcal{X}}$ having three slices (i.e., matrices). Let $\mathbf{X}^{(1)} \in \mathbb{R}_+^{F \times I}$ be the training speech utterance feature matrix, where F denotes the number of features and I is the number of training speech utterances. To capture the speaker's age, indicator vectors of dimension L are employed, where L is the number of levels employed to quantize the speaker age range. The speaker age matrix is denoted as $\mathbf{X}^{(2)} \in \mathbb{R}_+^{L \times I}$. Its li element

$X_{li}^{(2)}$ is 1 if the ith speaker falls into the domain of the lth quantization level and 0 otherwise. For example, let us consider $L = 10$ age intervals. The age intervals are carefully chosen in order to have an adequate (ideally, the same) number of observations in each interval and to cover the age range of all speakers of the dataset. Since we have only few utterances of speakers aged less than 28 years old or more than 84 years old, the first age interval represents speakers aged less than 28 and the last interval speakers aged more than 84. The 2nd to 9th age intervals have a range of 7 years. Then, a speaker aged 28 at the time of the recording is assigned to the 2nd age interval that corresponds to the age range [28-35), while an 83 years old speaker is assigned to the 9th age interval of range [77,84). Had the age intervals been less than 10, the orthogonality constraint imposed by PARAFAC2 (described in the next paragraph) would not be satisfied, while had the age intervals been more than 10, each age interval would contain very few observations, since the dataset is relatively small. Let us denote the third matrix as $\mathbf{X}^{(3)} \in \mathbb{R}_{+}^{M \times I}$, where M denotes the number of speakers. Its mi element $X_{mi}^{(3)}$ is 1 if the ith speech recording is uttered by the mth speaker. The speakers are grouped according to gender as follows. The first M_1 rows of matrix $\mathbf{X}^{(3)}$ are assigned to female speakers, while the remaining M_2 rows are assigned to male speakers. Clearly, $M_1 + M_2 = M$.

Since $\mathfrak{X}$ has three slices, the PARAFAC2 seeks a decomposition of the form:

$$\mathbf{X}^{(n)} = \mathbf{U}^{(n)}\, \mathbf{H}\, \mathbf{S}^{(n)}\, \mathbf{W}^T, \;\; n = 1, 2, 3 \tag{1}$$

where $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times k}$, $n = 1, 2, 3$ is an orthogonal matrix for each slice, $\mathbf{H} \in \mathbb{R}^{k \times k}$ is a square matrix, $\mathbf{S}^{(n)} \in \mathbb{R}^{k \times k}$ is a diagonal matrix of weights for the nth slice of $\mathfrak{X}$, and $\mathbf{W} \in \mathbb{R}^{I \times k}$ is a coefficient matrix. Clearly, $I_1 = F$, $I_2 = L$, and $I_3 = M$. Parameter k denotes the number of latent variables to be extracted from each utterance. To achieve uniqueness, the square matrix $(\mathbf{U}^{(n)}\mathbf{H})^T\, (\mathbf{U}^{(n)}\mathbf{H})$ is kept constant over n [20]. The decomposition (1) can be obtained by solving the optimization problem:

$$\underset{\mathbf{U}^{(n)},\, \mathbf{H},\, \mathbf{S}^{(n)},\, \mathbf{W}}{\operatorname{argmin}} \sum_{n=1}^{3} \|\mathbf{X}^{(n)} - \mathbf{U}^{(n)}\, \mathbf{H}\, \mathbf{S}^{(n)}\, \mathbf{W}^T\|_F^2. \tag{2}$$

The optimization problem (2) can be effectively solved with the algorithm described in [24]. Having solved the optimization problem (2), one computes the matrix $\mathbf{B} \triangleq \mathbf{U}^{(1)}\mathbf{H}\mathbf{S}^{(1)} \in \mathbb{R}_{+}^{F \times k}$. $\mathbf{B}$ spans a feature space of reduced dimensions k, where the semantic relations between the feature vectors and their associations with speaker's age and gender are retained. Indeed, the semantic relations between the age vectors as well as the gender vectors are propagated to the feature space through the common matrix of right singular vectors $\mathbf{W}$.

As long as the reduced dimensions feature space spanned by $\mathbf{B}$ is created, a test sketch is derived by pre-multiplying the feature vector extracted from an utterance $\mathbf{x} \in \mathbb{R}_{+}^{F \times 1}$ with $\mathbf{B}^{\dagger}$, i.e., $\tilde{\mathbf{x}} = \mathbf{B}^{\dagger}\, \mathbf{x} \in \mathbb{R}^{k \times 1}$. To predict the age interval of the speaker uttered the test utterance, one has to compute the vector $\mathbf{a} \in \mathbb{R}_{+}^{L \times 1}$ by

$$\mathbf{a} = \mathbf{U}^{(2)}\, \mathbf{H}\, \mathbf{S}^{(2)}\, \tilde{\mathbf{x}}. \tag{3}$$

The predicted age interval is associated with the largest value in $\mathbf{a}$. To predict the gender of the speaker who uttered the test utterance, one should compute the vector $\mathbf{g} \in \mathbb{R}_{+}^{M \times 1}$ given by

$$\mathbf{g} = \mathbf{U}^{(3)}\, \mathbf{H}\, \mathbf{S}^{(3)}\, \tilde{\mathbf{x}}. \tag{4}$$

If the largest value in $\mathbf{g}$ is located in the first M_1 elements of the vector, the speaker's gender is predicted to be female. Otherwise, a male speaker is predicted.

3. EXPERIMENTAL EVALUATION

3.1. Dataset

The longitudinal Trinity College Dublin Speaker Ageing (TCDSA) [12, 13] has been used in the experiments. The database contains recordings spanning a year range per speaker varying between 30 and 60 years at irregular intervals of between 1 to 10 years. The total number of speakers is 26, including 15 males and 11 females. The data were obtained from a variety of sources, such as television documentary series, YouTube, national broadcasters of U.K. and Ireland. Many different accents are included and there is a different number of recordings per speaker, varying from 4 to 47 recordings per speaker, compiling a total number of 280 recordings.

Furthermore, the duration of the recordings varies from 25 seconds to 35 minutes. In our experiments, a total duration of 30 seconds is kept from every recording. If the recording's duration is longer than 40 seconds, we discard the first 10 seconds and keep the following 30 seconds of the recording. If the recording's duration is shorter than 40 seconds, we keep the first 30 seconds of the recording or less if the recording lasts less than 30 seconds.

3.2. Auditory cortical representations

These feature descriptors are inspired by the way sound is perceived and processed by the human auditory system [21]. The human auditory system can be modeled by a two stage process. The first stage models the cochlea, and converts the audio signal to an auditory representation (spectrogram). Due to the fact that the basilar membrane across the cochlea exhibits a tonotopical organization, the basilar membrane can be modeled by a bank of bandpass filters. To this end, the constant Q transform (CQT) is employed [26]. The CQT is a technique, which transforms a signal from time to the frequency domain, such that the center frequencies of the bins are geometrically spaced and the Q factors (i.e., the ratios of the center frequencies to the bandwidths) are equal. This means that a better frequency resolution is observed for the low frequencies, while the time resolution is better for high frequencies, which resembles the frequency resolution of the auditory system.

In the second stage, the audio signal reaches the primary auditory cortex, where it is processed, perceived and interpreted. In this stage, the spectral and temporal modulation content of the auditory spectrogram is estimated. The cells in the primary auditory cortex are organized according to their response selectivity in different spectral and temporal stimuli [27]. To model this functionality, multi-resolution two-dimensional (2D) wavelet analysis is applied on the auditory spectrogram that was extracted in the first stage. The wavelet analysis is implemented using 2D Gaussian filters, ranging from narrow to broad spectral scales and from slow to fast temporal rates. The aforementioned analysis results in a four-dimensional (4D) representation of time, frequency, rate and scale, referred to as auditory cortical representation [21].

For the extraction of the auditory cortical representations, a number of parameters needs to be determined. Following [28], 128 filters were employed, which cover 8 octaves between 44.9 Hz and 11 kHz. Also, the elements of the CQT matrix were raised to the power of 0.1 in order to compress the magnitude of the CQT. Regarding the wavelet analysis of the second stage, a bank of 2D Gaussian filters was employed with scales $\in \{0.25, 0.5, 1, 2, 4, 8\}$ (Cycles/Octave) and rates $\in \{\pm 2, \pm 4, \pm 8, \pm 16, \pm 32\}$ (Hz). The resulting 4D representation was averaged on time and a 3D cortical representation (frequency, rate, and scale) was obtained. Subsequently, by re-arranging the elements of the 3D representation into a single vector, each utterance was described by a vector $\mathbf{x} \in \mathbb{R}_+^{F \times 1}$ for $F = 7680$ (i.e., 128 frequency channels × 10 rates × 6 scales).

3.3. Evaluation protocol and metrics

As mentioned before, the proposed method returns two ranking vectors for each test utterance. The first ranking vector is for predicting the speaker's age interval and the second one for predicting the speaker's gender. The latter prediction is a binary classification problem, while the former one is a multi-class classification process, where each age interval is considered as one class. Since we considered 10 age intervals, the number of classes is 10.

In order to assess the performance of the proposed framework in joint age and gender prediction, we conducted experiments on the TCDSA dataset. During the experimental evaluation, we applied 2-fold cross validation to the dataset consisting of 280 recordings. The number of folds was imposed by the small size of the dataset and the large number of age classes. In order to achieve a balanced training and test set in each fold, the recordings were assigned to train and test set by applying stratified sampling. Our goal was to include the same proportion of utterances in each age interval of the train and test set. To this end, we examined each age interval separately and the recordings in each interval were randomly partitioned into two halves. Half of the recordings in each age interval were used to build the train set, while the remaining ones built the test set. In the second fold, the roles of training and test set were reversed. The results disclosed in this paper refer to the mean of the evaluation metrics across the two folds.

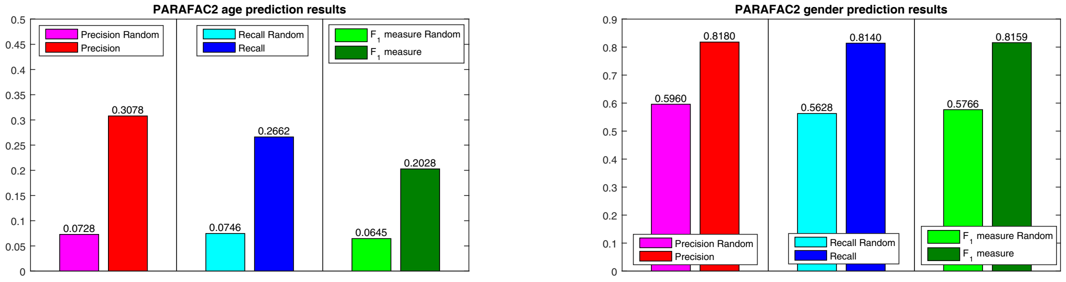
Precision, recall, and F_1 measure were employed as metrics to assess the predictions by the proposed method. We will briefly mention their definitions for age prediction. The definitions can be easily adapted for gender prediction. For each age class l among the $L = 10$ classes, the precision is the proportion of the test utterances predicted to belong to this age class by the proposed method that are correctly predicted to belong there. The recall is the proportion of the test utterances actually belonging to this age class that are correctly predicted to belong there. The F_1 measure is the averaged harmonic mean of precision and recall. Since age prediction is a multi-class classification problem, these metrics are calculated for each age class and micro-averaging is performed to yield a collective figure of merit.

3.4. Results

We applied PARAFAC2 with a number $k = 10$ of latent dimensions to the TCDSA dataset. The value of k was chosen so that the orthogonality constraint of PARAFAC2 is satisfied. In Figure 1, the mean values of evaluation metrics for $k = 10$ are presented for the PARAFAC2 and the Random model comparatively. As used in [29], the Random model gives a sense of the lowest expected value for each metric on a given dataset.

Let us describe the Random model for the gender prediction. Apparently, a similar procedure is applied for age prediction. The Random model samples the gender class (without replacement) from a multinomial distribution parameterized by the gender prior distribution, $P(i), i = 1, 2$ estimated using the observed gender in the training set [29]. Therefore, the gender selection according to the Random model relies on the gender appearance frequency, such that the most common gender is more likely to be chosen for a test utterance.

From the results depicted in Figure 1, we observe that the proposed method outperforms the random model in both tasks. Also, we notice that the evaluation metrics for gender prediction admit higher absolute values than those for age interval prediction. The better gender prediction results are not surprising, since predicting speaker's age from speech utterances is more difficult than predicting speaker's gender, even when the prediction is made by humans. To this end, we also investigate the predictions made by PARAFAC2 with some tolerance. More specifically, since our age intervals have a range of 7 years, if we consider as correct the age predictions



(a)

(b)

Fig. 1. PARAFAC2 prediction metrics for $k = 10$ latent dimensions on the TDCSA dataset used in [12, 13] against the same metrics for the Random model [29]: (a) Micro-averaged precision, micro-averaged recall, and micro-averaged F_1 measure in 2-fold cross validation for age interval prediction; (b) Mean precision, mean recall, and mean F_1 measure in 2-fold cross validation for gender prediction.

that differ only by one age class from the true age class, the predicted age intervals can be considered as correct with a tolerance of 7 years on average. For example, if a test utterance is predicted to belong to the 3nd age interval, while it truly belongs to the 2nd or the 4th age interval, then we can consider the prediction as approximately correct with a tolerance of 7 years on average. In Figure 2, the mean precision, the mean recall, and the mean F_1 measure for age prediction are plotted for $k = 10$ with and without tolerance in the predictions. The means are also derived from a 2-fold cross validation experiment.

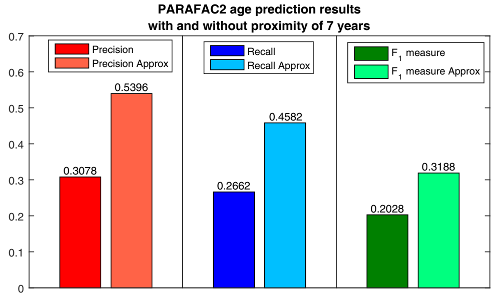


Fig. 2. PARAFAC2 model age prediction metrics for $k = 10$ on the TDCSA dataset used in [12, 13] with and without tolerance in age interval prediction.

4. CONCLUSIONS

An appealing automatic system for the prediction of speakers age and gender has been proposed. PARAFAC2 has been employed for semantically oriented feature extraction, age interval and gender prediction. The ranking scores returned by PARAFAC2 for age interval and gender prediction are used for multi-class and binary classification, respectively. The experimental results are promising and indicate the strength of PARAFAC2 to capture hidden relationships among the speech recordings.

The smallest values admitted by the figures of merit for age interval prediction than for gender prediction (Figure 1) challenge us to investigate alternative methods to build the speaker age matrix in the future. The critical aspect is to maintain the orthogonality of $\mathbf{U}^{(n)}$.

5. REFERENCES

[1] S. E. Linville, "The aging voice," *The American Speech-Language-Hearing Association Leader*, pp. 12–21, 2004.

[2] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4-93 years of age," *Journal of Speech, Language, and Hearing*, vol. 54, pp. 1011–1021, 2011.

[3] P. Torre III and J. A. Barlow, "Age-related changes in acoustic characteristics of adult speech," *Journal of Communication Disorders*, vol. 42, no. 5, pp. 324–333, 2009.

[4] F. Kelly and N. Harte, "Effects of long-term ageing on speaker verification," in *Biometrics and Identity Management BioID*, C. Vielhauer, J. Dittmann, A. Drygajlo, N.C. Juul, and M. Fairhurst, Eds., vol. 6583 of *LNCS*, pp. 113–124. Springer-Verlag, Heidelberg, 2011.

[5] Y. Matveev, "The problem of voice template aging in speaker recognition systems," in *Proc. SPECOM 2013*, M. Železný, I. Habernal, and A. Ronzhin, Eds., vol. 8113 of *LNAI*, pp. 345–353. Springer-Verlag, Heidelberg, 2013.

[6] A. Lanitis, "A survey of the effects of aging on biometric identity verification," *Int. Journal of Biometrics*, vol. 2, no. 1, pp. 34–52, 2010.

[7] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[8] 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. 8926, pp. 737–750, May 2015.

[9] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. Face and Gesture Recognition*, Southampton, UK, April 2006, pp. 341–345.

[10] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "Greybeard - voice and aging," in *Proc. Language Resources and Evaluation Conf.*, Malta, April 2010, pp. 2437–2440.

[11] J. D. Harnsberger, R. Shrivastav, and W. S. Brown Jr., "Modeling perceived vocal age in American English," in *Proc. Interspeech 2010*, Makuhari, Chiba, Japan, September 2010, pp. 466–469.

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

[13] F. Kelly and N. Harte, "Auditory detectability of vocal ageing and its effect on forensic automatic speaker recognition," in *Proc. Interspeech 2013*, Lyon, France, August 2013, pp. 2846–2850.

[14] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification in score-ageing-quality classification space," *Computer Speech & Language*, vol. 27, no. 5, pp. 1068–1084, 2012.

[15] F. Kelly, N. Bruümmer, and N. Harte, "Eigenaging compensation for speaker verification," in *Proc. Interspeech 2013*, Lyon, France, August 2013, pp. 1624–1628.

[16] F. Kelly, R. Saeidi, N. Harte, and D. van Leeuwen, "Effect of long-term ageing on i-vector speaker verification," in *Proc. Interspeech 2014*, Singapore, September 2014, pp. 86–90.

[17] C. Müller and F. Burkhardt, "Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age," in *Proc. Interspeech 2007*, Antwerp, Belgium, August 2007, pp. 2277–2280.

[18] F. Metze, J. Ajmera, R. Englert, F. Burkhardt, J. Stegmann, C. Müller, R. Huber, B. Andrassy, J. Bauer, and B. Littel, "Comparison of foure approaches to age and gender recognition for telephone applications," in *Proc. 2007 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, April 2007, vol. 4, pp. 1089–1092.

[19] W. Spiegl, G. Stemmer, E. Lasarcyk, V. Kolhatkar, A. Cassidy, B. Potard, S. Shum, Y. C. Song, P. Xu, P. Beyerlein, J. Harnsberger, and E. Nöth, "Analyzing features for automatic age estimation on cross-sectional data," in *Proc. Interspeech 2009*, Brighton, U.K., September 2009, pp. 2923–2926.

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

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

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

[23] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[24] P. Chew, B. Bader, T. Kolda, and A. Abdelali, "Cross-language information retrieval using PARAFAC2," in *Proc. 13th ACM Int. Conf. Knowledge Discovery and Data Mining*, San Jose, CA, USA, August 2007, pp. 143–152.

[25] Y. Panagakis and C. Kotropoulos, "Automatic music tagging via PARAFAC2," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 481–484.

[26] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. 7th Sound and Music Computing Conf.*, 2010.

[27] R. Munkong and B.-H. Juang, "Auditory perception and cognition," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 98–117, May 2008.

[28] 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 Processing*, vol. 22, no. 12, pp. 1905–1917, Dec. 2014.

[29] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, February 2008.