ARISTOTLE UNIVERSITY OF THESSALONIKI
FACULTY OF SCIENCES
DEPARTMENT OF INFORMATICS
POSTGRADUATE STUDIES PROGRAMME

Age interval and gender prediction using PARAFAC2 on speech recordings and face images

Pantraki Evangelia

Supervisor: Professor Kotropoulos Constantine

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

July 2016
I would like to dedicate this thesis to my beloved father.
Acknowledgements

My sincere gratitude goes to my postgraduate work supervisor, Professor Constantine Kotropoulos for his kind encouragement and patient supervision. It was a real privilege and an honour for me to share of his exceptional scientific knowledge but also of his extraordinary human qualities. His generous guidance and support made this thesis possible. Furthermore, his constructive remarks and careful editing contributed enormously to a complete presentation of this work.

Besides my supervisor, I would like to thank Associate Professor Andreas Lanitis in Cyprus University of Technology for his valuable suggestions, insightful comments and constructive criticisms at different stages of my research. Moreover, his hospitality has been exceptional and I feel obliged to express my gratitude.

Furthermore, I would like to thank my thesis committee members, Assistant Professor Nikos Laskaris and Assistant Professor Nikos Nikolaidis for their thoughtful comments and feedback. Moreover, a special thanks goes to all the members of the Artificial Intelligence and Information Analysis Laboratory of Department of Informatics at Aristotle University of Thessaloniki, as well as my fellow graduate students for their support.

Last but not least, I would particularly like to thank my parents, Emmanouil and Anastasia, and my sister Andromahi, for their continuous encouragement and support throughout this work. I am deeply grateful to family and friends, who have been an important and indispensable source of spiritual support.
Abstract

Parallel Factor Analysis (PARAFAC) is a generalization of Principal Component Analysis to the situation where a set of data matrices is to be analysed. Parallel Factor Analysis 2 (PARAFAC2) is a variant of PARAFAC that can be applied to unevenly sized multi-way arrays, i.e., data matrices having the same number of columns, but different numbers of rows. PARAFAC2 is powerful in extracting latent features and deriving a lower-dimensional semantic space, thus is ideally suited to classification problems. In this thesis, two novel soft biometric frameworks based on PARAFAC2 are proposed. The first performs gender and age interval prediction based on speech utterances, while the latter is bimodal and relies on both speech utterances and face images. In the unimodal soft biometric system, PARAFAC2 is applied to an irregular three-order tensor (or more precisely a hypermatrix) having three slices, while in the bimodal system PARAFAC2 is applied to an irregular fourth-order tensor. In each case, PARAFAC2 yields ranking vectors for age interval and gender prediction that are appropriately exploited for decision making. More specifically, in the bimodal system, Support Vector Machines (SVMs) are applied to the aforementioned ranking vectors. The Trinity College Dublin Speaker Ageing (TCDSA) [1, 2] dataset, that has been supplemented with face images of the speakers, has been used in the experiments. The experimental results indicate than the proposed methods constitute solid soft biometric verification techniques.
Περίληψη

Η παρούσα διπλωματική πραγματεύεται το πρόβλημα της πρόβλεψης του φύλου και της ηλικιακής ομάδας χρησιμοποιώντας την Παράλληλη Ανάλυση Παραγόντων 2 (Parallel Factor Analysis 2 - PARAFAC2). Η μέθοδος PARAFAC2 είναι μια μέθοδος αποσύνθεσης πινάκων, που αποτελεί γενίκευση της Ανάλυσης πίνακα σε ιδιάζουσες τιμές (Singular Value Decomposition) σε πολυδιάστατα δεδομένα. Η PARAFAC2 έχει τη δυνατότητα να πραγματοποιεί σημασιολογικά προσανατολισμένη μείωση διαστάσεων σε πολυδιάστατα δεδομένα με την εξαγωγή λανθάνουσων μεταβλητών. Μια σημαντική διαφορά της PARAFAC2 έναντι της PARAFAC, που αποτελεί και προπομπό της, είναι η δυνατότητα να εφαρμοστεί σε μη κανονικούς τένσορες (ή πιο σωστά υπερπίνακες). Ακόμη ένα πολύ σημαντικό χαρακτηριστικό της PARAFAC2 είναι η ικανότητά της να χειρίζεται δεδομένα που είναι επισήμευσιμά με πολλαπλές ετικέτες (labels) και να εξάγει έναν μικρότερων διαστάσεων σημασιολογικό χώρο που σχετίζεται με τις πολλαπλές ετικέτες.

Στην παρούσα διπλωματική, επιχειρείται μια καινοτόμα προσπάθεια αξιοποίησης των ιδιαίτερων χαρακτηριστικών της PARAFAC2 στο πεδίο της πρόβλεψης βιομετρικών χαρακτηριστικών (soft biometrics). Στόχος μας είναι να προβλέψουμε ταυτόχρονα την ηλικιακή ομάδα και το φύλο βασιζόμενοι στα βιομετρικά χαρακτηριστικά της ομιλίας και του προσώπου των ατόμων. Συγκεκριμένα, η προτεινόμενη μεθοδολογία εστιάζει στην αντιστάθμιση των επιδράσεων της γήρανσης (ageing) στην πρόβλεψη της ηλικιακής ομάδας και του φύλου. Η γήρανση επιδρά ποικιλοτρόπως στο πρόσωπο και τη φωνή των ατόμων και αποτελεί επιβαρυντικό παραγόντα για την απόδοση συστημάτων επαλήθευσης που
ηλικιακή κλάση των ομιλητών και τον πίνακα που φέρει πληροφορία για το γένος των ομιλητών. Η PARAFAC2 εφαρμόζεται στον προαναφερθέντα μη κανονικό τένσορα τριών διαστάσεων, ούτως ώστε οι σημασιολογικές συσχετίσεις μεταξύ των επισημεωσεων ηλικίας και φύλου των ηχογραφήσεων να οδηγήσουν στην εξαγωγή διανυσμάτων χαρακτηριστικών μειωμένων διαστάσεων που εμπεριέχουν σημασιολογική πληροφορία. Αυτό οφείλεται στην ιδιότητα της PARAFAC2 να αναπαριστά τα διανύσματα χαρακτηριστικών και τα αντίστοιχα διανύσματα της ηλικίας και του φύλου ως γραμμικούς συνδυασμούς διανυσμάτων βάσης με συντελεστές που προέρχονται από τον ίδιο διανυσματικό χώρο. Τα αριστερά ιδιάζοντα διανύσματα του πίνακα με τα χαρακτηριστικά της ηλικίας εκτείνονται σε έναν χαμηλότερων διαστάσεων σημασιολογικό χώρο που κυριαρχείται από την ηλικιακή και φυλετική πληροφορία. Κάθε διάνυσμα χαρακτηριστικών που αντιστοιχεί σε ένα δείγμα ομιλίας του συνόλου ελέγχου, αρχικά προβάλλεται σε αυτόν τον σημασιολογικό χώρο με στόχο να προκύψουν διανύσματα μικρότερων διαστάσεων. Για να προβλέψουμε την ηλικιακή ομάδα του ομιλητή στον οποίο ανήκει μια ηχογραφημένη ομιλία ελέγχου, το μειωμένων διαστάσεων διάνυσμα χαρακτηριστικών της συγκεκριμένης ομιλίας πολλαπλασιάζεται από τα αριστερά με τα αριστερά ιδιάζοντα διανύσματα του πίνακα που φέρει την ηλικιακή πληροφορία. Για να προβλεφθεί το φύλο του ομιλητή στον οποίο ανήκει μία ηχογραφημένη ομιλία ελέγχου, το μειωμένων διαστάσεων διάνυσμα χαρακτηριστικών της συγκεκριμένης ομιλίας πολλαπλασιάζεται από τα αριστερά με τα αριστερά
ιδιάζοντα διανύσματα του πίνακα που φέρει την φυλετική πληροφορία. Και στις δύο πε-ριπτώσεις, προκύπτει ένα διάνυσμα κατάταξης (ranking vector) το οποίο αξιοποιείται για την λήψη αποφάσεων. Τα πειραματικά δεδομένα από την αξιολόγηση της προτεινόμενης μεθόδου είναι ικανοποιητικά και καταδεικνύουν την δύναμη της PARAFAC2 να συλλάβει τις κρυμμένες συσχετίσεις μεταξύ των ηχογραφήσεων ομιλίας. Ειδικώτερα, η απόδοση της μεθόδου στην πρόβλεψη του φύλου παρουσιάζεται σαφώς ενισχυμένη σε σχέση με την απόδοση στην πρόβλεψη της ηλικιακής κλάσης.

Με στόχο μια πιθανή βελτίωση στην πρόβλεψη της ηλικιακής κλάσης, παρουσιάζεται ένα εναλλακτικό βιομετρικό σύστημα που βασίζεται σε δύο πηγές πληροφορίας για την πρόβλεψη της ηλικιακής ομάδας και του φύλου: σε δείγματα ομιλίας των ατόμων και σε εικόνες του προσώπου τους. Προφανώς, η ηλικία του ατόμου κατά τον χρόνο καταγραφής της ομιλίας και λήψης της εικόνας προσώπου είναι η ίδια. Ένα περιόδωρο ±3 χρόνων έγινε επιτρεπτό, προκειμένου να διευκολυνθεί η συλλογή των εικόνων που συμπλήρωσαν την αρχική βάση δεδομένων, η οποία αποτελούνταν μόνο από ηχογραφημένες ομιλίες και επίσης γιατί η επίδραση της γήρανσης σε διάστημα 3 χρόνων μπορεί να θεωρηθεί αμελητέα. Ένα σύστημα που βασίζεται σε δύο τύπους πληροφορίας είναι πιθανό να οδηγήσει σε βελτίωση της απόδοσης, καθώς μπορεί να εκμεταλλευτεί τη συμπλήρωματικότητα μεταξύ των δύο τύπων πληροφορίας. Εδώ, η PARAFAC2 εφαρμόζεται σε έναν τένσορα τεσσάρων διαστάσεων, αφού έχει συμπεριληφθεί και ο πίνακας με τα χαρακτηριστικά των εικόνων προσώπου. Ο τρόπος εφαρμογής της PARAFAC2 είναι όμοιος με το προηγούμε-νο πείραμα και όπως και σε εκείνο, προκύπτουν διανύσματα κατάταξης για την πρόβλεψη της ηλικιακής ομάδας και του φύλου. Συνολικά, η PARAFAC2 παρέχει δύο διανύσματα κατάταξης για την πρόβλεψη της ηλικιακής ομάδας, από τα οποία το ένα έχει προκύψει από τα δείγματα ομιλίας και ένα από τα δείγματα εικόνων προσώπου. Αντίστοιχα, η PARAFAC2 παρέχει δύο διανύσματα κατάταξης για την πρόβλεψη του φύλου, το καθένα από τα οποία βασίζεται σε έναν από τους δύο διαφορετικούς τύπους πληροφορίας. Στη
περιέχει, μια Μηχανή Εδραίων Διανυσμάτων (Support Vector Machine - SVM) εφαρμόζεται στα διανύσματα κατάταξης για την πρόβλεψη της ηλικιακής ομάδας. Αντίστοιχα, ένα SVM εφαρμόζεται στα αντίστοιχα διανύσματα κατάταξης για την πρόβλεψη του φύλου.

Τα δύο προτεινόμενα βιομετρικά συστήματα για την πρόβλεψη της ηλικιακής ομάδας και του φύλου συμψηφίζονται σε σε αρχή εν μέσω της PARAFAC2 για την σημασιολογικά προσανατολισμένη μείωση διαστάσεων των χαρακτηριστικών, αλλά η μεθοδολογία που ακολουθείται είναι διαφορετική σε χάρη σύστημα. Μια απευθείας σύγκριση στην απόδοση των δύο συστημάτων δεν είναι απολύτως δίκαιη, καθώς το σύνολο δεδομένων που χρησιμοποιείται στην πειραματική αξιολόγηση των δύο προτεινόμενων μεθόδων δεν περιλαμβάνει τον ίδιο αριθμό παρατηρήσεων. Όσον αφορά στην απόδοση του σύστημα που ενσωματώνει SVMs στην διαδικασία λήψης αποφάσεων είναι σαφώς πιο αποδοτικό από το σύστημα που δεν συμπεριλαμβάνει SVMs. Ειδικότερα, με την ενσωμάτωση
των SVMs στο βιομετρικό σύστημα, η τιμή της μετρικής αξιολόγησης $F_1$ measure για την πρόβλεψη της ηλικιακής ομάδας αυξήθηκε κατά 47.73% και κατά 63.05% όταν η πρόβλεψη γίνεται με μέση ανοχή 7 χρόνων στην πρόβλεψη της ηλικιακής κλάσης. Ιδιαίτερο ενδιαφέρον παρουσιάζει η βελτίωση που παρατηρείται στην πρόβλεψη της ηλικιακής ομάδας με τη συμπεριλήψη των εικόνων προσώπου στο μοντέλο της PARAFAC2. Συγκεκριμένα, ενώ η απόδοση του συστήματος που βασίζεται μόνο στις εικόνες προσώπου είναι χαμηλή σε σχέση με την απόδοση του συστήματος που βασίζεται μόνο στην ομιλία, η απόδοση του συστήματος ομιλίας βελτιώνεται ελαφρά με τη συμπεριλήψη της πληροφορίας των εικόνων προσώπου στο μοντέλο της PARAFAC2.

Όσον αφορά στην πρόβλεψη του φύλου, η απόδοση των προτεινόμενων μεθόδων είναι σαφώς ενισχυμένη σε σχέση με την πρόβλεψη της ηλικιακής ομάδας. Όπως και με την πρόβλεψη της ηλικιακής ομάδας, η απόδοση του προτεινόμενου συστήματος που βασίζεται στα αρχεία ομιλίας είναι σημαντικά καλύτερη από την απόδοση του συστήματος που βασίζεται στις εικόνες προσώπου. Ακόμα, η ενσωμάτωση των SVMs στο σύστημα δεν φαίνεται να επηρεάζει σημαντικά την απόδοση στην πρόβλεψη του φύλου, αντίθετα με ότι παρατηρήθηκε για την πρόβλεψη της ηλικιακής ομάδας.
Contents

Περίληψη ix

Contents xv

List of Figures xvii

List of Tables xix

1 Introduction 1
   1.1 General Background and Motivation . . . . . . . . . . . . . . . . . . . . . . . . 1
   1.2 Thesis Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
   1.3 Thesis Outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
   1.4 Notational Conventions . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2 Ageing within Biometric Verification Systems 5
   2.1 General Background . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
   2.2 Ageing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
   2.3 Applications . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

3 Parallel Factor Analysis 2 11
   3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
   3.2 Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3 Applications</td>
<td>16</td>
</tr>
<tr>
<td><strong>4 AGE INTERVAL AND GENDER PREDICTION USING PARAFAC2 APPLIED</strong></td>
<td></td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>19</td>
</tr>
<tr>
<td>4.2 Proposed method</td>
<td>20</td>
</tr>
<tr>
<td>4.3 Experimental Evaluation</td>
<td>23</td>
</tr>
<tr>
<td>4.3.1 Dataset</td>
<td>23</td>
</tr>
<tr>
<td>4.3.2 Auditory cortical representations</td>
<td>24</td>
</tr>
<tr>
<td>4.3.3 Evaluation protocol and metrics</td>
<td>25</td>
</tr>
<tr>
<td>4.3.4 Results</td>
<td>26</td>
</tr>
<tr>
<td>4.4 Conclusions</td>
<td>29</td>
</tr>
<tr>
<td><strong>5 BIMODAL AGE INTERVAL AND GENDER PREDICTION USING PARAFAC2 AND SVMS</strong></td>
<td></td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>31</td>
</tr>
<tr>
<td>5.2 Proposed method</td>
<td>32</td>
</tr>
<tr>
<td>5.3 Experimental Evaluation</td>
<td>36</td>
</tr>
<tr>
<td>5.3.1 Dataset</td>
<td>36</td>
</tr>
<tr>
<td>5.3.2 Feature extraction</td>
<td>38</td>
</tr>
<tr>
<td>5.3.3 Evaluation protocol and metrics</td>
<td>40</td>
</tr>
<tr>
<td>5.3.4 Results</td>
<td>44</td>
</tr>
<tr>
<td>5.3.5 Discussion</td>
<td>50</td>
</tr>
<tr>
<td>5.4 Conclusions</td>
<td>52</td>
</tr>
<tr>
<td><strong>6 Concluding Remarks and Future Extension</strong></td>
<td></td>
</tr>
<tr>
<td>Bibliography</td>
<td>57</td>
</tr>
</tbody>
</table>
List of Figures

4.1 PARAFAC2 prediction metrics for \( k = 10 \) latent dimensions on the TDCSA dataset used in [1, 2] against the same metrics for the Random model [3]:
(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. 27

4.2 PARAFAC2 model age prediction metrics for \( k = 10 \) on the TDCSA dataset used in [1, 2] with and without tolerance in age interval prediction. 28

5.1 Face images depicting four speakers of the TCDSA dataset at ascending ages. 38
5.2 Visualization of auditory cortical representations of speech samples of the same female person across ascending ages. 39
5.3 Visualization of auditory cortical representations of speech samples of the same male person across ascending ages. 39
5.4 Visualization of auditory cortical representations of speech samples of male and female persons that belong to the same age interval. 41
5.5 Distribution of selected values for parameter C of SVMs during the validation process: (a) for gender prediction; (b) for age interval prediction. 45
5.6 Face images with wrong gender predicted via PARAFAC2+SVM. 48
5.7 Face images with erroneous age interval predicted via PARAFAC2+SVM.
The persons depicted in the first row (3 face images) were predicted 1 age interval younger and in the second row 1 age interval older. . . . . . . . . 49
## List of Tables

5.1 $F_1$ measure results for gender prediction derived from the proposed method. Comparatively, results for PARAFAC2 with 3 slices and results with 4 slices that are based solely on speech and on image features are presented. 47

5.2 $F_1$ measure results for age prediction derived from the proposed method. Comparatively, results for PARAFAC2 with 3 slices and results with 4 slices that are based solely on speech and on image features are presented. The approximate results for age prediction refer to classification results with one age class tolerance. 47

5.3 Average weighted score from the questionnaire on age prediction. 50

5.4 $F_1$ measure results for age prediction derived from the unimodal system presented in Section 4, a PARAFAC2+SVM system where PARAFAC2 is trained on speech feature vectors (3 slices) and SVM on speech ranking vectors, a PARAFAC2+SVM system where PARAFAC2 is trained on speech and face image feature vectors (4 slices) and SVM on speech ranking vectors and a PARAFAC2+SVM system where PARAFAC2 is trained on speech and face image feature vectors (4 slices) and SVMs on speech and image ranking vectors, respectively. 51
5.5 $F_1$ measure results for gender prediction derived from the unimodal system presented in Section 4, a PARAFAC2+SVM system where PARAFAC2 is trained on speech feature vectors (3 slices) and SVM on speech ranking vectors, a PARAFAC2+SVM system where PARAFAC2 is trained on speech and face image feature vectors (4 slices) and SVM on speech ranking vectors and a PARAFAC2+SVM system where PARAFAC2 is trained on speech and face image feature vectors (4 slices) and SVMs on speech and face image ranking vectors, respectively. . . . . . . . . . . . . . . . . . . . . . . . . . 52
Chapter 1

Introduction

1.1 General Background and Motivation

Parallel Factor Analysis (PARAFAC) is a multi-way generalization of the Singular Value Decomposition (SVD) [4] that can be used for determining explanatory factors for third or higher-order data. While PARAFAC can be only applied to regular tensors, a variant of PARAFAC, namely Parallel Factor Analysis 2 (PARAFAC2), can be applied to irregular multidimensional matrices. A procedure for fitting the PARAFAC2 model directly to a set of data matrices is proposed in [5].

PARAFAC2 is able to perform dimensionality reduction to multi-way data by extracting latent variables. One of the first applications of PARAFAC2 involved modelling chromatographic data with retention time shifts [6]. Another important characteristic of PARAFAC2 is its ability to handle multi-labelled data and derive a lower-dimensional semantic space related to multiple labels. Therefore, it is suitable for multi-label classification problems. It has been applied to multi-label classification of documents [7], music recordings [8], and images [9].

Here, a contribution to the field of soft biometrics relying on PARAFAC2 is investigated. Biometric templates are snapshots of biometric characteristics captured at a particular time
In a verification scenario, a user is first enrolled and attempts to access the system, if he or she is verified positively, later on. The decision is influenced heavily whenever a time lapse between the enrolment and the verification exists [11]. Ageing is a factor that influences greatly both speaker verification [12] and face verification [13]. Great research effort has focused on establishing speaker verification techniques able to handle the diverse effects of ageing [1][2][14][15]. Furthermore, the automatic prediction of biometric characteristics based on facial characteristics [16][17] has attracted great research interest.

While many unimodal biometric verification systems based solely on one biometric characteristic, e.g., vocal [1], facial [18], fingerprint [19] characteristics, have been developed, very few multi-modal verification systems have been proposed. One of the most recent attempts is reported in [20], where a multi-modal verification system with vocal and visual biometric modalities is presented. The same modalities are utilized in [21] for user verification and demonstrate improved performance compared to the corresponding unimodal systems.

In this thesis, an effort to exploit the distinctive properties of PARAFAC2 in order to perform age interval and gender prediction is proposed. Our goal is to jointly predict persons’ age interval and gender based on their speech utterances or both speech utterances and face images. Firstly, a unimodal biometric system is developed for the classification of gender and age interval, which is based solely on speech utterances across a wide age range of speakers’ life. Subsequently, in order to exploit the complementary information and improve prediction performance, a bimodal verification system is proposed that takes into account both speech utterances and face images in order to predict genre and age intervals. Both approaches, unimodal and bimodal, utilize the powerful decomposition properties of PARAFAC2, but differ in the procedure followed in order to make predictions.
1.2 Thesis Contributions

The motivation for this thesis stems from the undisputed strength of PARAFAC2 to perform semantically oriented dimensionality reduction. Due to its power in extracting latent variables while combining multiple sources of information, PARAFAC2 emerges as a novel and promising technique to address multi-classification problems. The problem of age interval and gender prediction is addressed, including both speech and facial verification. A quite intriguing task is to combine different types of information, which may be complementary to each other, in order to boost performance. In summary, the contributions of this work are:

1. The novelty of employing PARAFAC2 to perform prediction of genre and age interval.

2. The proposal of a unimodal system employing PARAFAC2 in order to predict persons’ age interval and gender based on their speech utterances.

3. The proposal of a bimodal system that exploits two modalities of information: speech utterances and face images. The proposed framework utilizes PARAFAC2 and Support Vector Machines (SVMs) in order to jointly predict the age interval and gender.

1.3 Thesis Outline

The remainder of this thesis is structured as follows: A review of the current research activity in ageing within biometric verification systems is presented in Section 2. In particular, the state-of-the-art techniques for the prediction of biometric characteristics, with emphasis on gender and age estimation, are thoroughly investigated. Section 3 analyzes the fundamentals of PARAFAC2 and its theoretical and mathematical background. Additionally, many applications of PARAFAC2 across different research areas are cited and briefly described. A novel unimodal speaker verification system that is based on speech utterances
is proposed in Section 4. The proposed framework that utilizes vocal features in order to predict gender and age interval is presented in detail in Section 4. Experimental results of the proposed framework applied to the Trinity College Dublin Speaker Ageing (TCDSA) Database [1, 2] are presented in Section 4. In Section 5, a bimodal system is presented that predicts the biometric characteristics of interest based on both speech utterances and face images. The bimodal framework relies on PARAFAC2 too, but incorporates SVMs in order to make predictions. More specifically, SVMs are applied to the predictive ranking vectors derived via PARAFAC2. Finally, Section 6 concludes the thesis and recommends directions for future work.

1.4 Notational Conventions

For the sake of consistency, throughout this thesis the following notation is adopted. Let \( \mathbb{Z} \) and \( \mathbb{R} \) denote the set of integer and real numbers, respectively. A tensor is a multidimensional array. The order of a tensor is the number of its dimensions, also known as ways or modes. More formally, an \( n \)-way or \( n \)th-order tensor is an element of the tensor product of \( n \) vector spaces, each of which has its own coordinate system [22]. Tensors of order three or higher are denoted by boldface Euler script calligraphic letters (e.g., \( \mathcal{X} \)), tensors of order two, i.e., matrices, are denoted by uppercase boldface letters (e.g., \( \mathbf{U} \)), tensors of order one, i.e., vectors, are denoted by lowercase boldface letters (e.g., \( \mathbf{u} \)) and scalars are denoted by lowercase letters (e.g., \( u \)). A third-order real-valued tensor \( \mathcal{X} \) 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 \( \mathcal{X} \) is addressed by 3 indices, i.e., \( x_{i_1 i_2 i_3} \). A fourth-order real-valued tensor \( \mathcal{X} \) is defined over the tensor space \( \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n} \), where \( I_1, \ldots, I_n \in \mathbb{Z} \) and \( n = 1, 2, \ldots, 4 \). A fourth-order tensor \( \mathcal{X} \) has four indices and each element of \( \mathcal{X} \) is addressed by \( x_{i_1 i_2 i_3 i_4} \). Hereafter, the operations on tensors are expressed in matricized form [22]. Throughout the thesis, the Frobenius matrix norm is denoted by \( \| F \) and the Moore-Penrose pseudoinverse of \( \mathbf{B} \) is denoted by \( \mathbf{B}^\dagger \).
Chapter 2

Ageing within Biometric Verification Systems

2.1 General Background

The term biometrics comes from the ancient Greek words bios = “life” and metron = “measure”. A biometric characteristic is defined as a measurable, physical characteristic or personal behavioural trait used to recognize the identity, or verify the claimed identity, of an individual [23]. Facial images, fingerprints, hand geometry, retina geometry, signature, and iris scan samples are all examples of biometrics [24].

Biometric templates are actually snapshots of biometric characteristics captured at a particular time instant [10]. The operation of a typical biometric verification system involves two stages: the enrollment stage and the verification stage. During enrolment, a user’s biometric data is acquired, a feature vector that represents the acquired data is extracted and finally, the feature vectors are stored in the database. The feature set of a specific user is the biometric template of the user. Since some biometric characteristics, especially behavioural biometrics such as voice and signature, present great variability, multiple templates for each user should be stored. The second stage of a biometric verification system is the verification
stage. Here, a user attempts to access the system and provides biometric data. Features are extracted from the biometric data and are, consequently, compared to the templates already stored in the database. If there is a match, the user’s identity is validated and the user is granted access to the system [25]. If a time lapse between the enrolment and the verification exists, the verification decision is heavily influenced [11]. Especially physiological biometric characteristics, such as vocal and facial characteristics, are notably altered with the pass of time due to the effects of ageing.

### 2.2 Ageing

Ageing, i.e., the process of growing old, has various effects on human face and voice. The acoustic changes of the human voice as a result of ageing, known as *vocal ageing*, have been thoroughly studied in [26–28]. 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 [12].

The changes of the human face as a result of ageing, known as *facial ageing*, are described in detail in [29, 30]. Ageing causes changes in both hard and soft facial tissues, such as the skeletal structure, skin, facial musculature and lines, forehead, lips, and ears. Each part is affected individually by the ageing process, but all-together lead to facial senescence. Interestingly, the ageing effects on face images can be simulated in order to predict how a young person might look in the future or how an old person might have looked in the past [31].
Since ageing causes changes to biometric characteristics, it is a factor that introduces undesirable variability into a biometric verification system. Ageing can be an important issue when it comes to speaker recognition [32], as well as face recognition [31], since it can deteriorate the ability of humans and machines to identify aged individuals.

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 [33] or the MORPH database [34], 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 [35], the University of Florida Vocal Aging Database [36], and the longitudinal Trinity College Dublin Speaker Ageing (TCDSA) database [1].

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 [1]. The addition of temporal information to the mel frequency cepstral coefficients (MFCCs) caused an increase in the rate of degradation [12]. However, at time-lapse of 30 years, vocal ageing caused significant problems in forensic automatic speaker recognition [2]. 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 [37]. 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 [38]. An age estimation system
based on i-vectors is also presented in [39]. Firstly, Linear Discriminant Analysis (LDA) was performed to reduce the dimension of i-vectors and subsequently, Support Vector Regression (SVR) models are utilized for automatic age estimation. PLS-Ranker, a Partial Least Squares based Ranker for age estimation is proposed in [40]. PLS-Ranker employs a coding strategy in order to estimate age based on the decision of many binary classifiers. The threshold of each binary classifier is determined via an adaptive threshold learning technique. Three novel systems combining short-term cepstral features and long-term features for speaker age recognition were compared to each other in [15]. A system combining GMMs using frame-based MFCCs and Support Vector Machines (SVMs) 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 [41]. 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 [42]. A structured-prediction based mechanism for the automatic computation of Voice Onset Time (VOT) and Voice Offset Time (VOFT) is presented in [43] and their predictive power for age estimation is examined.

Many methods have been developed for the automatic prediction of biometric characteristics based on facial characteristics and the extraction of facial ageing patterns [16]. A survey on soft biometrics identification methods based on facial features is presented in [17]. The survey reveals a common framework used to address different soft biometric issues that involves feature extraction from face images, exploitation of subspace dimensionality reduction techniques and feeding the image features of reduced dimension into classifiers, most frequently SVM classifiers due to their efficiency in computational cost and accuracy. Another survey for computer-based age synthesis and estimation via facial appearance is presented in [18]. Classification for age interval prediction is also demon-
strated in [44], where 5 age classes are considered, Principle Component Analysis (PCA) is adopted to reduce the feature dimensions and a SVM with Gaussian Radian Basis Function (RBF) is utilized for decision making. The joint prediction of age and gender from facial characteristics has attracted great research activity. In [45], Convolutional Neural Networks were deployed for the task of joint age interval and gender classification, while SVMs were deployed in [46].

While many approaches for age interval and gender prediction based on either speech or facial characteristics have been reported, research activity has also been made in the area of multi-modal verification systems during the past decade. One of the first attempts is reported in [20], where a multi-modal verification system with vocal and visual biometric modalities is presented. The same modalities are utilized in [21] for user verification and demonstrate improved performance compared to the corresponding uni-modal systems. In [47], information from three biometric modalities, namely face, fingerprint, and hand geometry, is combined at the matching score level in order to perform biometric user verification. An interesting approach for information fusion of different modalities based on SVMs for identity verification is described in [48]. In [49], a multi-modal verification system based on speech and face images is presented. The system is adaptive, i.e., it adapts to the noise conditions of the speech samples.
Chapter 3

Parallel Factor Analysis 2

3.1 Introduction

PARAFAC can be considered as a multi-way generalization of the SVD. It is a multi-way decomposition method that can be used for determining explanatory factors for third or higher-order data [4]. PARAFAC2 [50] 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 of matrices forms the slices of an irregular third-order tensor. In other words, while PARAFAC can only be applied to regular tensors, PARAFAC2 can be applied to irregular tensors, i.e., collections of matrices with only one common dimension.

A procedure for fitting the PARAFAC2 model directly to a set of data matrices is proposed in [5]. Like the PARAFAC model, PARAFAC2 also provides unique solutions under specific constraints [50]. 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.

### 3.2 Overview

Let us consider an irregular tensor $\mathcal{X}$ having the slices $X_k \in M_k \times N, \ k = 1, 2, \ldots, K$. The application of the PARAFAC2 model on each slice of the hypermatrix $\mathcal{X}$ has the following form:

$$X_k = U_k H S_k V^T$$

In Equation (3.1), matrix $U_k$ is an orthonormal $M_k \times R$ factor matrix for each slice of $\mathcal{X}$, where $R$ is the number of latent variables extracted via PARAFAC2. Additionally, $H \in R \times R$ is a square matrix, $S_k$ is an $R \times R$ diagonal matrix of weights for the $k$th slice of $\mathcal{X}$, and $V \in N \times R$ is a factor matrix. In order to gain uniqueness, Harshman [4, 50] imposed the constraint that the cross product $(U_k H)^T (U_k H) = C$, where $C \in R \times R$ is constant over $k$. Since $U_k$ is orthonormal, this is equivalent to $H^T (U_k^T U_k) H = C \Rightarrow H^T H = C$, i.e., $H^T H$ is constant over $k$. So, the imposed constraint is accomplished when $H$ is non-singular. In PARAFAC2 model, different orthonormal factor matrices $U_k$ are allowed for each slice of tensor $\mathcal{X}$, while in PARAFAC the factor matrix $U$ is constrained to be the same across slices.

To compute the PARAFAC2 model of tensor $\mathcal{X}$, an algorithm is presented in [7] whose complete procedure is as follows:

1. Initialize $V = R$, where $R$ holds the $R$ principal components of $\sum_k X_k^T X_k \in N \times N$.
   The matrices $H$ and $S_1, S_2, \ldots, S_K$ are initialized as $R \times R$ identity matrices.

2. Compute SVD of $Z_k = H S_k V^T X_k = P_k \sum_k Q_k^T$ by first computing the $R$ principal eigenvectors of $Z_k Z_k^T \in R \times R$ to obtain $P_k$ and normalizing the columns of $Z_k^T P_k$ to obtain $Q_k \in M_k \times M_k$ and then update $U_k = Q_k P_k^T$, $k = 1, 2, \ldots, K$.

3. Update $H, V, S_1, \ldots, S_K$ by one iteration of an Alternating Least Squares algorithm.
for the standard PARAFAC applied to the $R \times N \times K$ tensor with slices $U_k^T X_k, k = 1, ..., K$.

4. Repeat Steps 2-3 until a maximum number of iterations has been reached or the norm of the residual $\sum_k \|X_k - U_k H S_k V^T\|$ ceases to change.

The aforementioned procedure of computing the PARAFAC2 mode, presented in [7], is followed in this thesis. It is a variant of the direct fitting algorithm for the PARAFAC2 model outlined in [5]. In [5], the PARAFAC2 model is defined as

$$ X_k = F_k D_k A^T + R_k $$  \hspace{1cm} (3.2)

In Equation (3.2), matrix $X_k \in I_k \times J$ is the $k$ slice of tensor $\mathcal{X}$, $F_k \in I_k \times R$ is a matrix of factor scores, $D_k$ is a diagonal $R \times R$ matrix containing the weights for the $k$ slice of $\mathcal{X}$, $A \in J \times R$ is a matrix of weights for the column units, and $R_k$ denotes the $I_k \times J$ matrix with residuals, $k = 1, ..., K$. The only invariance constraint imposed in (3.2) is that the loading matrices $A D_1, ..., A D_k$ are proportional. The aforementioned invariance constraint is not sufficient to achieve uniqueness. To this end, Harshman [4] imposed a constraint on the factor scores, i.e., the constraint that the cross product matrix $F_k^T F_k = F$ is constant over $k$. In [5], it is proposed to fit the model (3.2) itself to the data by minimizing:

$$ \sigma_1(F_k, A, D_1, ..., D_k) = \sum_{k=1}^{K} \|X_k - F_k D_k A^T\|^2 $$

subject to $F_k^T F_k = F_l^T F_l$ for all pairs $k, l = 1, ..., k$

$$ D_k \text{ is diagonal matrix, } k = 1, ..., K $$  \hspace{1cm} (3.3)

In [5], it is proven that the first constraint is equivalent to the constraint $F_k = P_k F$, where $P_k$ is a column orthonormal $n_k \times R$ matrix and $F$ is an $R \times R$ matrix. So the problem of
minimizing (3.3) can be expressed as minimizing

$$
\sigma_2(P_1, ..., P_K, F, A, D_1, ..., D_K) = \sum_{k=1}^{K} \|X_k - P_k F D_k A^T\|^2
$$

subject to $P_k^T P_k = I_R$

(3.4)

$D_k$ is diagonal matrix, $k = 1, ..., K$

In order to minimize (3.4), an Alternating Least Squares algorithm is proposed in [5] that alternately minimizes (3.4) over $P_k$ for fixed $F$, $D_k$ and $A$, $k = 1, ..., K$, and over $F$, $D_1, ..., D_k$ and $A$ for fixed $P_1, ..., P_k$.

Let us attempt to minimize $P_1$ over for fixed $F$, $D_k$ and $A$, $k = 1, ..., K$. The minimization objective is to minimize:

$$
\|X_1 - P_1 F D_1 A^T\|^2
$$

subject to $P_1^T P_1 = I_R$

(3.5)

$D_1$ is diagonal matrix

According to Golub and Van Loan [51], the minimization problem of (3.5) can be rewritten as:

$$
\|X_1 - P_1 F D_1 A^T\|^2 = \text{tr}[(X_1 - P_1 F D_1 A^T)(X_1 - P_1 F D_1 A^T)^T] = \\
\text{tr}[(X_1 - P_1 F D_1 A^T)(X_1^T - A D_1 F^T P_1^T)] = \\
\text{tr}(X_1 X_1^T - P_1 F D_1 A^T X_1^T - X_1 A D_1 F^T P_1^T + P_1 F D_1 A^T A D_1 F^T P_1^T) = \\
\text{tr}(X_1 X_1^T - 2P_1 F D_1 A^T X_1^T - X_1 A D_1 F^T P_1^T + P_1 F D_1 A^T A D_1 F^T P_1^T)^{\text{fixed}}
$$

(3.6)

The last term of equation (3.6) is fixed since $F$, $D_1$ and $A$ are fixed. Thus, in order to
minimize (3.5), it suffices to maximize:

\[ f(P_1) = \text{tr}(P_1 F D_1 A^T X_1^T) = \text{tr}(F D_1 A^T X_1^T P_1) \]

subject to \( P_1^T P_1 = I_R \)

\[ D_1 \text{ is diagonal matrix} \]

The maximization problem (3.7) can be rewritten as:

\[ f(P_1) = \text{tr}(P_1^T (F D_1 A^T X_1^T)^T) = \text{tr}(P_1^T X_1 A D_1 F^T) \]

(3.8)

According to [51], the maximizing of \( P_1 \) can be found by calculating the SVD of \( X_1 A D_1 F^T \). Let us denote the aforementioned SVD as \( U_1 \Sigma_1 V_1^T \). It is proven than the maximum of (3.8) over columnwise orthonormal \( P_1 \) is given by \( P_1 = V_1 U_1^T \). The proof is as follows.

Let us define the orthogonal matrix \( Z_1 = V_1^T P_1^T U_1 \). Then, (3.8) can be written as:

\[ f(P_1) = \text{tr}[P_1^T (X_1 A D_1 F^T)] = \text{tr}(P_1^T U_1 \Sigma_1 V_1^T) = \text{tr}\left(\frac{V_1^T P_1^T U_1 \Sigma_1}{Z_1}\right) = \]

\[ \text{tr}(Z_1 \Sigma_1) = \sum_{i=1}^{R} z_{ii} \sigma_i \leq \sum_{i=1}^{R} \sigma_i \]

(3.9)

Clearly, the upperbound can be found by setting \( P_1 = U_1 V_1^T \), since with substitution we have the identity matrix:

\[ Z_1 = V_1^T P_1^T U_1 = V_1^T (U_1 V_1^T)^T U_1 = V_1^T V_1 U_1^T U_1 = I_R \]

(3.10)

Similarly for each slice, \( P_k \) is maximized for \( P_k = U_k V_k^T = V_k U_k^T \) [5]. So, the problem
of minimizing (3.4) over $F, D_1 \ldots D_K$ and $A$ reduces to minimizing:

$$\sigma_2(F, A, D_1 \ldots D_K | P_1 \ldots P_K) = \sum_{k=1}^{K} \|P_k^T X_k - F D_k A^T\|^2 + c$$  \hspace{1cm} (3.11)

where $c$ denotes a constant with respect to $F, A, D_1 \ldots D_k$.

The minimization problem (3.11) can be minimized over $F, D_1 \ldots D_k$ and $A$ using any PARAFAC algorithm. Instead of minimizing (3.11) over $F, D_1 \ldots D_k$ and $A$, it suffices to decrease (3.11) over the parameter matrices, as is achieved by using one cycle of updates from a PARAFAC algorithm.

### 3.3 Applications

A wide range of PARAFAC2 applications to chemometrics has been reported. In [6], PARAFAC2 was employed in order to understand the chemistry of the color formation during sugar processing from beets. PARAFAC2 is able to model the retention time shifts of chromatographic data. Similarly, PARAFAC2 combined with PCA is used in chromatographic analysis in [52]. Also in the chemometrics field, PARAFAC2 has been applied to fault detection and diagnosis in semiconductor etch [53].

PARAFAC2 was employed with novelty for the simultaneous detection and classification of neural action potentials [54]. The method was applied to Micro Electrode Array (MEA) data, whose structure is multidimensional. Since different source types (e.g., cells, which respond directly to external stimuli and cells, which are triggered by other neural cells) are characterized by different spike shapes, PARAFAC2 is employed in order to separate the spike shapes (sources) in time, frequency and space (channels).

PARAFAC2 has been applied successfully to feature extraction and multi-label classification of documents [7]. More specifically, in [7] PARAFAC2 was used for cross-language information retrieval, since it can delineate between different languages unlike Latent Se-
mantic Analysis (LSA). In LSA, all languages have to be concatenated together for training in order to create a multilingual “document”. Instead, PARAFAC2 provides the ability to create a parallel corpus. Let $X_k$ be the term-by-document matrix for the $k$th language in the parallel corpus. PARAFAC2 is applied to every matrix $X_k$ (i.e., for every different language):

$$X_k = U_k H S_k V^T \quad (3.12)$$

The matrix $V$ of right singular “document” vectors is the same across all languages. To gain uniqueness, the constraint that the cross-product $(U_k H)^T (U_k H)$ is constant over $k$ [4, 50]. In other words, PARAFAC2 imposes the constraint, not present in standard LSA, that the “concepts” (i.e., columns of $U_k$) of any given language in the parallel corpus have a separate mapping into the LSA conceptual space. Intuitively, this constraint makes sense, since the whole purpose of using a parallel corpus is that translations are supposed to render the same concepts in different languages. During experimental evaluation on parallel corpus of Bible, PARAFAC2 emerged as a superior alternative to standard LSA for multilingual information retrieval [7].

Additionally, the powerful properties of PARAFAC2 were exploited for music tagging [8]. Auditory temporal modulations were extracted from each music recording and a third-order tensor was composed that had two slices: the vectorized training temporal modulations and the corresponding multi-label vectors. PARAFAC2 was applied to this tensor in order to perform semantically oriented feature extraction and multi-label annotation of music recordings.

Moreover, PARAFAC2 was employed for automatic image tagging and recommendation of images to users of a social network [9]. PARAFAC2 was applied to a collection of three matrices, namely visual features, user information regarding each image and tag information for each image. Experimental evaluation indicated that the inclusion of three slices in the PARAFAC2 model enabled the capture of the latent relations between the images features,
the tags, and the user interests demonstrating promising results.

In [55], SCREAM (Shifted Covariates REgression Analysis for Multi-way data), a novel method for multi-way regression problems with shifts and shape changes in one mode is presented. SCREAM is based on a combination of PARAFAC2 and principal covariates regression (PCovR). The method has been evaluated on different simulated and real datasets and yielded satisfactory results.

Moreover, in [56] a novel speaker adaptation based on PARAFAC2 of transformation matrices for continuous speech recognition is presented. An acoustic model adaptation method is presented, where the transformation matrix for a new speaker is given by the product of bases and a weight matrix. The bases are built from PARAFAC2 of training speakers’ transformation matrices. The utilization of PARAFAC2 to the continuous speech recognition problem proves to be an appealing approach.

An application of PARAFAC2 to gait analysis is reported in [57]. The study of locomotion involves analyzing multi-waveform data (e.g., velocities, accelerations, etc.) from various body locations (e.g., knees, ankles, etc.) of several subjects. PARAFAC2 was applied for the detection and analysis of individual and group differences in multi-joint multi-waveform gait data.
Chapter 4

AGE INTERVAL AND GENDER PREDICTION USING PARAFAC2 APPLIED TO SPEECH UTTERANCES

4.1 Introduction

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 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
A novel framework for age and gender prediction is proposed that is based on PARAFAC2 [50]. In the training phase, the starting point is to form an irregular third-order tensor (or more precisely a 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 [58] 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 informa-
4.2 Proposed method

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 longitudinal Trinity College Dublin Speaker Ageing (TCDSA) [1, 2] database, using a 2-fold cross validation protocol.

A PARAFAC2 model is trained on an irregular third-order tensor $X$ having three slices (i.e., matrices). Let $X^{(1)} \in \mathbb{R}^{F \times I}_+$ 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 $X^{(2)} \in \mathbb{R}^{L \times I}_+$. Its $li$ element $X^{(2)}_{li}$ is 1 if the $i$th speaker falls into the domain of the $l$th 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 \( X^{(3)} \in \mathbb{R}^{M \times I} \), where \( M \) denotes the number of speakers. Its \( m_i \) element \( X_{mi}^{(3)} \) is 1 if the \( i \)th speech recording is uttered by the \( m \)th speaker. The speakers are grouped according to gender as follows. The first \( M_1 \) rows of matrix \( X^{(3)} \) are assigned to female speakers, while the remaining \( M_2 \) rows are assigned to male speakers. Clearly, \( M_1 + M_2 = M \).

Since \( X \) has three slices, the PARAFAC2 seeks a decomposition of the form:

\[
X^{(n)} = U^{(n)} H S^{(n)} W^T, \quad n = 1, 2, 3
\]  

(4.1)

where \( U^{(n)} \in \mathbb{R}^{I_n \times k}, n = 1, 2, 3 \) is an orthogonal matrix for each slice, \( H \in \mathbb{R}^{k \times k} \) is a square matrix, \( S^{(n)} \in \mathbb{R}^{k \times k} \) is a diagonal matrix of weights for the \( n \)th slice of \( X \), and \( W \in \mathbb{R}^{I \times k} \) 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 \( (U^{(n)}H)^T (U^{(n)}H) \) is kept constant over \( n \) [50]. The decomposition (4.1) can be obtained by solving the optimization problem:

\[
\arg\min_{U^{(n)}, H, S^{(n)}, W} \sum_{n=1}^{3} \left\| X^{(n)} - U^{(n)} H S^{(n)} W^T \right\|_F^2.
\]  

subject to \( (U^{(n)})^T U^{(n)} = I \)  

(4.2)

The optimization problem (4.2) can be effectively solved with the algorithm described in [7]. Having solved the optimization problem (4.2), one computes the matrix \( B \triangleq U^{(1)} H S^{(1)} \in \mathbb{R}^{F \times k} \). \( B \) 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 \( W \).

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

\[
\mathbf{a} = \mathbf{U}^{(2)} \mathbf{HS}^{(2)} \hat{\mathbf{x}}.
\] (4.3)

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

\[
\mathbf{g} = \mathbf{U}^{(3)} \mathbf{HS}^{(3)} \hat{\mathbf{x}}.
\] (4.4)

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.

### 4.3 Experimental Evaluation

#### 4.3.1 Dataset

The longitudinal TCDSA database [1, 2] 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.

### 4.3.2 Auditory cortical representations

These feature descriptors are inspired by the way sound is perceived and processed by the human auditory system [58]. 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 [59]. 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 [60]. 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 [58].

For the extraction of the auditory cortical representations, a number of parameters needs to be determined. Following [61], 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 $\mathbf{x} \in \mathbb{R}_{\pm}^{F \times 1}$ for $F = 7680$ (i.e., 128 frequency channels $\times$ 10 rates $\times$ 6 scales).

4.3.3 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 Section 4.3.4 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.

### 4.3.4 Results

We applied PARAFAC2 with a number of $k = 10$ latent dimensions to the TCDSA dataset. The value of $k$ was chosen so that the orthogonality constraint of PARAFAC2 is satisfied. In Figure 4.1, the mean values of evaluation metrics for $k = 10$ are presented for the PARAFAC2 and the Random model comparatively. As used in [3], 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 [3]. 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 4.1, we observe that the proposed method outperforms the Random model in both tasks. Also, we notice that the evaluation metrics for gen-
Figure 4.1 PARAFAC2 prediction metrics for $k = 10$ latent dimensions on the TDCSA dataset used in [1, 2] against the same metrics for the Random model [3]: (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.

der 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 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 3rd 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 4.2, 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.

Figure 4.2 PARAFAC2 model age prediction metrics for $k = 10$ on the TDCSA dataset used in [1, 2] with and without tolerance in age interval prediction.
4.4 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 4.1) challenge us to investigate alternative methods to build the speaker age matrix in the future. The critical aspect is to maintain the orthogonality of $U^{(n)}$. 
Chapter 5

BIMODAL AGE INTERVAL AND GENDER PREDICTION USING PARAFAC2 AND SVMS

5.1 Introduction

The prediction of a person’s age and gender is a useful and challenging task in biometrics/forensics. In this section, bimodal age interval and gender prediction is proposed that is based on both visual and aural features. The expectation from a bimodal system is performance improvement by exploiting any complementarity of the aforementioned information sources. PARAFAC2 is deployed due to its power to efficiently model multi-way data and extract latent variables. PARAFAC2 reduces the dimensions of visual and aural features and provides ranking vectors for the prediction of age interval and gender. Subsequently, an SVM classifier is applied to the ranking vectors derived from PARAFAC2 for conducting predictions. The aforementioned procedure is applied to the Trinity College Dublin Speaker Ageing database [1, 2] that is supplemented with face images of the speakers. These face images are captured at an age close to the speaker’s age allowing for ±3 year tolerance. Ex-
Experimental results indicate solid gender and promising age interval prediction performance.

5.2 Proposed method

In Section 4 [62], we proposed a uni-modal speech based automatic age interval and gender prediction system based on PARAFAC2 [50]. Here, a bimodal framework is proposed that resorts to both audio and visual features in order to predict age interval and gender. The proposed bimodal framework also utilizes PARAFAC2 to extract latent features and perform dimensionality reduction, but is complemented by SVMs which are applied to the ranking vectors derived via PARAFAC2. In the training phase, the starting point is to form an irregular fourth-order tensor (or more precisely a hypermatrix) having four 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 [58] and their derivation is motivated by the human auditory system. The second slice is the face image feature matrix, whose columns are the face image raw intensity values. The third slice is the speaker age matrix whose columns are indicator vectors of suitable dimension associated to speaker’s age intervals. The fourth 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 fourth-order tensor so that the semantic similarities between the age and gender annotations of the utterances and face images 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 vectors and the associated age intervals 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 and the
5.2 Proposed method

face image feature matrix span semantic spaces of lower dimensionality dominated by the age and gender information.

Any auditory cortical representation vector extracted from a test utterance and accordingly, any face image feature vector extracted from a test face image, are first projected onto their corresponding semantic space in order to obtain the test sketches. To predict speaker’s age interval associated to a test utterance and a face image, the test speech sketch and the test face image are pre-multiplied by the matrices arisen from the SVD of the speaker age matrix. Similarly, to predict the gender of the speaker who uttered any test recording and is depicted on any test face image, the test speech sketch and the test face image sketch are pre-multiplied by the matrices arisen from the SVD of the speaker gender matrix. In both cases, two ranking vectors are derived that are exploited for decision making. Promising results are demonstrated when the aforementioned framework is applied to the TCDSA database that is supplemented with contemporary face images of each speaker, using a Leave-One-Out Cross-Validation validation protocol.

A PARAFAC2 model is trained on an irregular fourth-order tensor \( \mathbf{X} \) having four slices (i.e., matrices). Let \( \mathbf{X}^{(1)} \in \mathbb{R}^{F_1 \times I'} \) be the training speech utterance feature matrix, where \( F_1 \) denotes the number of audio features and \( I' \) is the number of training speech utterances. Similarly, let \( \mathbf{X}^{(2)} \in \mathbb{R}^{F_2 \times I''} \) be the training face image feature matrix, where \( F_2 \) denotes the number of image features and \( I'' \) is the number of training face images. Each speech utterance is matched with a face image allowing for a tolerance of \( \pm 3 \) years of its capturing date from speaker’s age. To represent the person’s age, indicator vectors of dimension \( L \) are employed, where \( L \) is the number of levels persons’ age is quantized to. The age matrix is denoted as \( \mathbf{X}^{(3)} \in \mathbb{R}^{L \times I'} \). Its \( li \) element \( X^{(3)}_{li} \) is 1 if the \( i \)th person’s age falls into the domain of the \( l \)th 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 entire age range of all
speakers in a dataset. Since in our case we have only few utterances of speakers aged less than 25 years old or more than 81 years old, the first age interval represents speakers aged less than 25 and the last interval speakers aged greater than 81. The 2nd to 9th age intervals have a range of 7 years. Then, a speaker aged 25 at the time of the recording is assigned to the 2nd age interval that corresponds to the age range [25-32), while an 80 years old speaker is assigned to the 9th age interval of range [74,81). 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. Let us denote the fourth matrix as $X^{(4)} \in \mathbb{R}^{M \times I_{tr} +}$, where $M$ denotes the number of persons. Its $m_i$ element $X^{(4)}_{m_i}$ is 1 if the $i$th speech recording is uttered by the $m$th speaker and likewise $i$th face image belongs to $m$th person, too. The persons are grouped according to gender as follows. The first $M_1 = 11$ rows of matrix $X^{(4)}$ are assigned to female persons, while the remaining $M_2 = 14$ rows are assigned to male persons. Clearly, $M = M_1 + M_2 = 25$ speakers.

Since $X$ has four slices, the PARAFAC2 seeks a decomposition of the form:

$$X^{(n)} = U^{(n)} H S^{(n)} W^T, \quad n = 1, 2, \ldots, 4$$

(5.1)

where $U^{(n)} \in \mathbb{R}^{I_n \times k}, \ n = 1, 2, \ldots, 4$ is an orthogonal matrix for each slice, $H \in \mathbb{R}^{k \times k}$ is a square matrix, $S^{(n)} \in \mathbb{R}^{k \times k}$ is a diagonal matrix of weights for the $n$th slice of $X$, and $W \in \mathbb{R}^{I_{tr} \times k}$ is a coefficient matrix. Clearly, $I_1 = F_1$, $I_2 = F_2$, $I_3 = L$, and $I_4 = M$. Parameter $k$ denotes the number of latent variables to be extracted from each utterance and face image, respectively. To achieve uniqueness, the square matrix $(U^{(n)} H)^T (U^{(n)} H)$ is kept constant over $n$ [50]. The decomposition (5.1) subject to the aforementioned orthogonality constraints for $U^{(n)}$ can be obtained by solving the optimization problem:

$$\arg\min_{U^{(n)}, H, S^{(n)}, W} \sum_{n=1}^{4} \|X^{(n)} - U^{(n)} H S^{(n)} W^T\|_F^2.$$  (5.2)
The optimization problem (5.2) can be effectively solved with the algorithm described in [7]. Having solved the optimization problem (5.2), one computes the matrix $B_1 \triangleq U^{(1)} H S^{(1)} \in \mathbb{R}^{F_1 \times k}$. $B_1$ spans a speech feature space of dimension $k$, where the semantic relations between the speech feature vectors and their associations with speaker’s face image features, age, and gender are retained. Similarly, $B_2 \triangleq U^{(2)} H S^{(2)} \in \mathbb{R}^{F_2 \times k}$ spans a face image feature space of reduced dimension $k$, where the semantic relations between the image feature vectors and their associations with person’s speech features, age, and gender are retained. Indeed, the semantic relations between the age vectors as well as the gender vectors are propagated to the feature spaces through the common matrix of right singular vectors $W$.

Having derived the speech and the face image feature spaces of reduced dimensions (spanned by $B_1$ and $B_2$, respectively), we proceed to a validation stage aiming to tune the parameters of a classifier applied to validation sketches, i.e., reduced dimension feature vectors, in order to predict the genre and the age interval. During validation, for each audio feature vector $x_{v1}^i$, a sketch $\tilde{x}_{v1}^i$ is derived by pre-multiplying the feature vector $x_{v1}^i \in \mathbb{R}^{F_1 \times 1}$ with $B_1^\dagger$, i.e., $\tilde{x}_{v1}^i = B_1^\dagger x_{v1}^i \in \mathbb{R}^{k \times 1}$. Similarly, for each face image feature vector $x_{v2}^i$, another sketch $\tilde{x}_{v2}^i = B_2^\dagger x_{v2}^i \in \mathbb{R}^{k \times 1}$ is computed. Needless to say that both $\tilde{x}_{v1}^i$ and $\tilde{x}_{v2}^i$ bear information from all slices through the bottleneck model matrix $H$, which is present in both $B_1$ and $B_2$.

Next, ranking vectors for age interval and gender prediction are derived. In particular, the ranking vector for age interval prediction from validation speech sketch $\tilde{x}_{v1}^i$ is obtained as

$$a_{v1}^i = U^{(3)} H S^{(3)} \tilde{x}_{v1}^i. \quad (5.3)$$

Likewise, the ranking vector for age interval prediction from validation face sketch $\tilde{x}_{v2}^i$ is found as

$$a_{v2}^i = U^{(3)} H S^{(3)} \tilde{x}_{v2}^i. \quad (5.4)$$

By concatenating the two age ranking vectors $a_{v1}^i \in \mathbb{R}^{L \times 1}$ and $a_{v2}^i \in \mathbb{R}^{L \times 1}$, the augmented ranking vector $a^v = [a_{v1}^T | a_{v2}^T]^T \in \mathbb{R}^{2L \times 1}$ is formed. Let us denote by $A^v \in \mathbb{R}^{2L \times I}$ the matrix
whose columns are the age ranking vectors for all validation measurements. Subsequently, an SVM employing a linear kernel is trained for age interval estimation. The SVM is fed by the columns of $A^v$.

A similar procedure is followed for gender prediction. Starting from a ranking vector for gender prediction from validation speech sketch $\tilde{x}_1^v$ and face sketch $\tilde{x}_2^v$

$$g_i^v = U^{(4)} HS^{(4)} \tilde{x}_i^v, \ i = 1, 2.$$  

(5.5)

the augmented ranking vector $g^v = [g_1^v T | g_2^v T]^T \in \mathbb{R}^{2M \times 1}$ is formed. A second SVM employing a linear kernel is trained for gender prediction. This SVM is applied to the columns of matrix $G^v$ associated to the ranking vectors of all validation measurements.

During the test phase, first the sketches from a test speech utterance and its associated test face image are computed and then the augmented ranking vectors $a^{te} = [a_1^{te} T | a_2^{te} T]^T$ and $g^{te} = [g_1^{te} T | g_2^{te} T]^T$ are derived. The trained SVM is applied to $a^{te}$ for age interval prediction. Gender prediction is obtained by the second trained SVM, which is fed by $g^{te}$.

### 5.3 Experimental Evaluation

#### 5.3.1 Dataset

The longitudinal TCDSA database [1, 2] has been used in the experiments, as in Section 4.3.1. 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 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. The total number of speakers included in the TCDSA dataset is 26, including 15 males and 11
In order to conduct bimodal age interval and gender prediction, face images were collected for each speaker of the dataset. Effort has been devoted so that face images were captured close to the speakers’ age. Since the exact matching was difficult, a 3 year tolerance was allowed between the age of a person when his/her face was captured and the age associated to his/her utterance. That is, if a person was 24 years old when his/her utterance was recorded, face images of this person were sought when he/she aged between 21 and 27.

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 each recording. If the recording duration is longer than 40 seconds, we discard the first 10 seconds and retain the following 30 seconds. If the recording duration is shorter than 40 seconds, we retain the first 30 seconds or less if the recording lasts less than 30 seconds. If many face images of the person at the age of the speech recording are collected and the recording has a duration longer than 40 seconds, more than one segments of 30 seconds long are kept. Of course, the number of the segments a speech recording is split into depends on the number of contemporary face images for the same person found and the overall duration of the speech recording. For example, if a recording of speaker aged 42 has a total duration of 198 seconds (3 minutes and 18 seconds), and 3 images of the speaker at the ages of 42, 45 and 45, respectively have been collected, the recording is split into 3 segments of 30 seconds each.

The original TCDSA dataset includes 280 recordings prior to splitting into segments. A total of 236 face images for the TCDSA speakers were gathered. After the age matching, a total of 227 recordings with contemporary face images were retained. No age matches were found between recordings and face images for one male speaker in the database. So, finally the total number of speakers included in the extended TCDSA audio-visual dataset was 25, including 14 males and 11 females.

The face images have size $140 \times 180$ pixels and depict approximately the same face
area. Furthermore, attention has been paid so that the images have the same depth of field. The photo shot type is Close up, where the head of the depicted person is visible from the top of the person’s hair till the top of the shoulder. Is is also called a “head shot” [63]. Nevertheless, the pose may vary. Also, the illumination varies greatly over the collected face images. Some examples of the collected face images for four speakers of the TCDSA dataset are depicted in Figure 5.1.

Figure 5.1 Face images depicting four speakers of the TCDSA dataset at ascending ages.

5.3.2 Feature extraction

For each speech recording and face image, a feature vector is extracted as follows. As in Section 4.3.2, auditory cortical representations are extracted from speech. These descriptors are inspired by the way sound is perceived and processed by the human auditory system [58]. For their extraction, a number of parameters needs to be determined. Following [61], 128 filters are employed, which cover 8 octaves between 44.9 Hz and 11 kHz. Each utterance is described by a vector $x_1 \in \mathbb{R}^{F_1 \times 1}$ where $F_1 = 7680$ (i.e., 128 frequency channels × 10
5.3 Experimental Evaluation

rates × 6 scales). In Figure 5.2, a visualization of the auditory cortical representations for a female speaker across ascending ages is depicted. Accordingly, a visualization of the auditory cortical representations for a male speaker across ascending ages is depicted in Figure 5.3. Furthermore, in Figure 5.4, visualizations of the auditory cortical representations of males and females belonging to the same age interval are depicted.

Figure 5.2 Visualization of auditory cortical representations of speech samples of the same female person across ascending ages.

Figure 5.3 Visualization of auditory cortical representations of speech samples of the same male person across ascending ages.
The image features consist of the pixel intensity values. Firstly, all images are converted to gray scale and a Discrete Cosine Transform (DCT)-based normalization is applied [64]. The DCT-normalization removes some of the low-frequency information that is susceptible to illumination changes. As mentioned before, the size of each image is 140 × 180 pixels, so each vectorized image has size $F_2 = 25200$.

### 5.3.3 Evaluation protocol and metrics

In order to assess the performance of the proposed framework in bimodal age interval and gender prediction, we conducted experiments on the extended TCDSA dataset including contemporary speakers’ face images. The extended dataset comprises $I = 227$ observations. During the evaluation, Leave-One-Out Cross-Validation (LOOCV) protocol is applied. In a LOOCV protocol, the number of folds equals the number of observations. Successively, each observation is considered as test set while the remaining $I - 1$ observations are used for training and validation.

As described before, the proposed method yields ranking vectors $g \in \mathbb{R}^{2M \times 1}$ for gender prediction, where $M$ equals the number of persons in the dataset and ranking vectors $a \in \mathbb{R}^{2L \times 1}$ for age interval prediction, where $L$ equals the number of age intervals. In order to make predictions, SVM classifiers are applied to these ranking vectors. More specifically, a first SVM is applied to the gender ranking vectors in order to predict gender and a second SVM is applied to the age ranking vectors in order to predict an age interval. The prediction of gender constitutes a binary classification problem, while the prediction of age interval is treated as a multi-class classification problem with 10 age classes, since we considered 10 different age intervals.

For running SVMs, we used the LIBSVM package [65]. The type of classifier is C-Support Vector Classification with a linear kernel. The best value for parameter $C$, i.e., the cost parameter of SVM, is selected based on the performance in the validation set. More
Figure 5.4 Visualization of auditory cortical representations of speech samples of male and female persons that belong to the same age interval.
specifically, in each fold of LOOCV one observation serves as test sample, 20% of the 226 observations (i.e., 45) are exploited for validation, and the remaining observations (i.e., 181) compose the train set. In order to achieve a balanced training and validation set in each fold, the recordings were assigned to train and validation sets by applying stratified sampling. Our goal was to include the same proportion of observations of each age interval in the train and validation set. To this end, at each fold, we examined each age interval separately and the observations belonging to each interval were randomly partitioned by 80% into the train set and 20% into the validation set.

A range of different values for parameter $C$ was examined for both SVMs; the one trained on gender ranking vectors and the one trained on age ranking vectors. The value of parameter $C$ that yields the best result on the validation test is used for training the SVM and subsequently, predicting gender and the age interval of the test observation. For age interval prediction, the “one-against-one” approach is followed, which is default option in the LIBSVM package and proved as most suitable multi-class approach in [66].

Precision, recall, and $F_1$ measure were employed as metrics to assess the predictions made by the proposed method. Their definitions for age prediction are briefly mentioned next. 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 measurements correctly predicted to belong to this age class by the proposed method. The recall is the proportion of the test measurements 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. Micro-averaging pools per-measurement decisions across classes, and then computes the evaluation metrics on the pooled contingency table. Micro-averaging gives equal weight to each per-measurement classification decision. Because the $F_1$ measure ignores true nega-
tives and its magnitude is mostly determined by the number of true positives, large classes dominate small classes in micro-averaging [67]. However, since age classes are balanced, the utilization of micro-averaging results in equal values for precision and recall. Since precision and recall are equal and $F_1$ measure is their averaged harmonic mean, $F_1$ measure is also equal to precision and recall. Thus, it suffices to employ only one evaluation metric, namely the $F_1$ measure.

Moreover, the performance of the proposed method on age interval prediction was compared to human performance on predicting age based on speakers’ utterances and face images. To this end, a questionnaire\(^1\) was designed that was composed of 24 questions. The respondent was presented with pairs of samples of speech recordings, face images, and both speech recordings and face images of the same person and was asked to determine which sample belonged to the person at an older age. The questionnaire comprised of two parts. In Part 1 (12 questions), 3 questions were asked about 4 sample persons of the dataset. Here, the questions regarding each person were presented successively, i.e., the first question for each person presented pairs of speech utterances, the second question presented pairs of face images and the third question pairs of both speech utterances and face images of the same person, respectively. Likewise, the following 3 questions presented the same series of information regarding the next sample person. In Part 2 (12 questions), 3 questions were asked about 4 sample persons of the dataset, but the order of the presented questions was different from Part 1. The first 4 questions presented only pairs of speech utterances at different age intervals for the 4 sample persons, the following 4 questions presented pairs of face images, and the last 4 questions presented both speech utterances and face images of the sample persons.

Throughout the questionnaire, each question included 6 possible answers. For each pair of samples of speech, face images, combined speech and face images, let us denote the first part of the pair as A and the second part as B. The 6 possible answers were: “A is

\(^{1}\text{http://testageestimation.polldaddy.com/s/test-of-age-estimation}
definitely older”, “A is probably older”, “Same age”, “B is probably older”, “B is definitely older”, “Not sure”. In order to evaluate the answers, different weights were assigned to each answer, i.e., if person A is actually older than person B, the weights assigned to the 5 possible answers are 5, 4, 3, 2, 1, 0, respectively. Correspondingly, if person B is actually older than person A, the weights assigned to the 5 possible answers are considered at reverse order, i.e., 1, 2, 3, 4, 5, 0. Apparently, at each case, the answer “Not sure” has zero weight.

5.3.4 Results

The extended TCDSA dataset described in Section 5.3.1 was used in the experiments. Firstly, PARAFAC2 yielded augmented ranking vectors for gender and age interval prediction by jointly processing speech utterances and contemporary face images. A number of $k = 10$ latent dimensions were extracted via PARAFAC2 for each of the 4 slices. The value of $k$ was chosen, so that the orthogonality constraint required for $U_n, n = 1, 2, \ldots, 4$ is satisfied. Secondly, the augmented ranking vectors $g^{te}$ and $a^{te}$ were fed to the dedicated SVM for either gender or age interval prediction using the LOOCV protocol detailed in Section 5.3.3. LOOCV defines 227 folds. In each fold, a grid searching was performed for determining the value of $C$ that yielded the top $F_1$ measure for each prediction task in the validation dataset associated to the fold. The histograms of the selected values for parameter $C$ during validation across the 227 folds for either gender or age interval prediction are depicted in Figure 5.5. It is seen that the most frequent top performing value of $C$ for gender prediction was $2^{-5}$. For age interval prediction, the most frequent top performing value of $C$ was 0.5.

In Tables 5.1 and 5.2, the micro-averaged $F_1$ measure for PARAFAC2 with $k = 10$ and the Random model are presented. The latter model gives a sense of the lowest expected value for the metric under consideration on a given dataset. Let us describe the Random model for gender prediction [3]. Apparently, a similar procedure was applied to age inter-
Figure 5.5 Distribution of selected values for parameter C of SVMs during the validation process: (a) for gender prediction; (b) for age interval prediction.

The Random model samples the gender class (without replacement) from a multinomial distribution parameterized by the gender prior distribution $P(i), i = 1, 2$. The gender prior distribution is estimated using the observed gender in the training set [3]. Accordingly, the gender selection by the Random model relies on the gender appearance fre-
quency. That is, the most frequent gender in the training set is more likely to be chosen for a test observation.

The $F_1$ measure for gender prediction in several experiments is summarized in Table 5.1. More specifically, an irregular third order tensor of 3 slices, namely the speech feature, the age interval, and gender indicator matrices, was decomposed by PARAFAC2 in the second column. The gender was predicted by a linear SVM applied to the gender ranking vectors of the test utterances. Likewise, an irregular third order tensor of 3 slices, namely the image feature, the age interval, and gender indicator matrices, was decomposed by PARAFAC2 in the third column and an SVM was applied to the test gender ranking vectors derived via PARAFAC2. Subsequently, the same exercise was conducted by a fourth order irregular tensor of four slices, namely the speech feature, the image feature, the age interval and the gender indicator matrices, which was decomposed by PARAFAC2 in the fourth and fifth column of Table 5.1. However, gender prediction was made by an SVM trained only on $g_1^r$ ranking vectors and tested on $g_1^r$ ranking vectors in the fourth column. Similarly, gender prediction was made by an SVM trained only on $g_2^r$ ranking vectors and tested on the $g_2^r$ ranking vectors in the fifth column. Finally, bimodal gender prediction by applying the method described in Section 5.2 to the augmented ranking vectors was assessed in the sixth column of Table 5.1.

It is seen that the proposed method that employs PARAFAC2 to extract ranking vectors that are subsequently classified by an SVM outperformed the Random model. Furthermore, gender prediction based on speech was found to be more accurate than that based on face image features. The aforementioned observation was found to be true in all experiments conducted. The inclusion of image features yielded a small performance degradation (4th column and 6th column). Reading Table 5.1 from the point of view of face image features, predictions based solely on face image were drastically improved (i.e., increase in $F_1$ measure by 0.22) when speech features and the associated gender ranking vectors were included.
5.3 Experimental Evaluation

Some examples of face images where gender was misclassified by PARAFAC2+SVM are presented in Figure 5.6.

<table>
<thead>
<tr>
<th>Gender prediction results</th>
<th>Speech, 3 Slices</th>
<th>Image, 3 Slices</th>
<th>Speech, 4 Slices</th>
<th>Image, 4 Slices</th>
<th>Speech+Image, 4 Slices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random model</td>
<td>0.5198</td>
<td>0.5154</td>
<td>0.4978</td>
<td>0.4846</td>
<td>0.5198</td>
</tr>
<tr>
<td>PARAFAC2 + SVM</td>
<td>0.8370</td>
<td>0.6079</td>
<td>0.8018</td>
<td>0.5551</td>
<td>0.8282</td>
</tr>
</tbody>
</table>

Table 5.1 $F_1$ measure results for gender prediction derived from the proposed method. Comparatively, results for PARAFAC2 with 3 slices and results with 4 slices that are based solely on speech and on image features are presented.

<table>
<thead>
<tr>
<th>Age prediction results</th>
<th>Speech, 3 Slices</th>
<th>Image, 3 Slices</th>
<th>Speech, 4 Slices</th>
<th>Image, 4 Slices</th>
<th>Speech+Image, 4 Slices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random model</td>
<td>0.0969</td>
<td>0.0793</td>
<td>0.1233</td>
<td>0.0837</td>
<td>0.1101</td>
</tr>
<tr>
<td>PARAFAC2 + SVM</td>
<td>0.2996</td>
<td>0.1145</td>
<td>0.3040</td>
<td>0.1101</td>
<td>0.2247</td>
</tr>
<tr>
<td>Approx Random</td>
<td>0.2952</td>
<td>0.3128</td>
<td>0.3480</td>
<td>0.2511</td>
<td>0.2996</td>
</tr>
<tr>
<td>Approx PARAFAC2 + SVM</td>
<td>0.5198</td>
<td>0.4273</td>
<td>0.5683</td>
<td>0.4141</td>
<td>0.5374</td>
</tr>
</tbody>
</table>

Table 5.2 $F_1$ measure results for age prediction derived from the proposed method. Comparatively, results for PARAFAC2 with 3 slices and results with 4 slices that are based solely on speech and on image features are presented. The approximate results for age prediction refer to classification results with one age class tolerance.

Table 5.2 presents age interval prediction conducting six experiments as in Table 5.1. In addition, we investigated also the case when some tolerance in age interval prediction was allowed. More specifically, since the age intervals in the dataset have a range of 7 years, we consider as correct any predictions that differ by one age interval from the actual age interval. That is, the predicted age intervals are treated as correct with a tolerance of 7 years on average. For example, if a test observation is predicted to belong to the 3rd age interval, while it truly belongs to the 2nd or the 4th age interval, then we consider the prediction as approximately correct within the tolerance allowed. The $F_1$ measure for age interval prediction when tolerance is allowed for either Random or PARAFAC2+SVM are gathered in the last two rows of Table 5.2.

The Figures of merit collected in Table 5.2 demonstrate a great performance improvement when tolerance on age interval prediction is allowed for both Random model and the combination of PARAFAC2 and SVM. The proposed PARAFAC2+SVM method always
outperforms the Random model. The $F_1$ measure admitted by age interval predictions based on speech features is more accurate than those based on face image features. The best $F_1$ measure of 0.3040 for PARAFAC2+SVM was obtained by classifying the ranking vectors driven by sketch speech features when the latter were derived by employing face image features in PARAFAC2 decomposition. However, the performance gain is marginal (i.e., $4.4 \times 10^{-3}$). By classifying the augmented ranking vectors with a linear SVM, the $F_1$ measure drops by 0.0793 (4th and 6th column). When tolerance in age interval prediction is allowed, the top $F_1$ measure of 0.5683 was measured for PARAFAC2+SVM that classifies ranking vectors driven by sketch speech features when the latter were derived by employing face image features in PARAFAC2 decomposition. The performance gain against predictions based on speech features exclusively is worth noticing (i.e., 0.0485). When the predictions are based on augmented ranking vectors driven by both speech and face image features, the $F_1$ measure is reduced by 0.0309, but still is greater than the prediction based on speech features exclusively by 0.0176. Some examples of face images where age interval was misclassified by PARAFAC2+SVM are presented in Figure 5.7.

![Figure 5.6 Face images with wrong gender predicted via PARAFAC2+SVM.](image)

To test whether the evaluation metric differences between the PARAFAC2+SVM and the baseline Random model are statistically significant, we apply the probabilistic approach presented in [68]. To this end, the probability distributions of $F_1$ measure for PARAFAC2+SVM and Random model are obtained and compared. The probability that PARAFAC2+SVM
achieves a better $F_1$ measure score than the Random model for gender prediction reaches 100%, 97.34%, 100%, 92.89% and 100% for each of the experiments presented in Table 5.1, respectively. For age prediction, PARAFAC2+SVM outperforms the Random model with probability of 100%, 98.56%, 100%, 83.59% and 99.98% for each of the experiments presented in Table 5.2, respectively. Similarly, when an average of 7 years tolerance is allowed, the difference between the $F_1$ measure of PARAFAC2+SVM and the Random model is highly significant, since the probability that the proposed method is better than Random model reaches 100%, 99.56%, 100%, 99.96% and 100% for each of the experiments summarized at Table 5.2, respectively.

Finally, the performance of humans in age prediction was assessed using the questionnaire detailed in Section 5.3.3. In total 51 persons answered the questionnaire. 49% of the respondents were female and 51% were male. The vast majority of the respondents were aged 21-30 (76%). The second largest group comprised 14% of the respondents whose age was between 31 and 40 years old. As described in Section 5.3.3, the questionnaire included two parts, which differed in the order of the presented samples of each modality. Each question had five possible answers and the average weighted score was calculated. The higher the score, the better the performance. The scores gathered are summarized in Table 5.3. The score was higher when the age prediction was based on face images than when it was
BIMODAL AGE INTERVAL AND GENDER PREDICTION USING PARAFAC2 AND SVMS

based on speech in both parts of the questionnaire. The humans seem to predict the age more accurately when they are looking at the face images than when they are listening to speech utterances. The just mentioned observation does not match the performance of the proposed method which admits the highest $F_1$ measure when speech features are employed.

<table>
<thead>
<tr>
<th>Scores speech</th>
<th>Scores images</th>
<th>Scores speech+image</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>std</td>
<td>mean</td>
</tr>
<tr>
<td>Part A</td>
<td>3.215</td>
<td>0.3</td>
</tr>
<tr>
<td>Part B</td>
<td>2.95</td>
<td>0.2</td>
</tr>
<tr>
<td>Total</td>
<td>3.0825</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 5.3 Average weighted score from the questionnaire on age prediction.

5.3.5 Discussion

The common ground between the unimodal approach presented in Section 4.2 and the bimodal approach presented in this chapter (Section 5.2) is the utilization of PARAFAC2 for semantically oriented extraction of latent variables. Nevertheless, the methodology proposed here is rather different than the one used in the unimodal approach. Here, SVMs are employed in order to contribute to the decision making process.

A straightforward compare between the systems’ performance is not completely fair, since the dataset used in the experimental evaluation of each framework is not entirely the same and doesn’t include the same number of samples. More specifically, during the completion of the TCDSA dataset, which initially included only speech samples, with contemporary face images, a complete matching between the recordings with contemporary face images could not be achieved. Nevertheless, a comparative analysis between the two proposed methods will be attempted, taking into account the aforementioned considerations.

In Table 5.4, the age prediction experimental results of the two proposed methods are comparatively presented. The results summarized on Table 5.4 have been already presented on Figure 4.2 of Section 5.3.5 and Table 5.2 of this chapter. The inclusion of SVMs in the
proposed framework leads to an improvement in the $F_1$ measure results of age prediction. More specifically, the $F_1$ measure for age prediction (0.2028) where no SVM was applied increased by 47.73% (0.2996) and by 63.05% (0.5198) for approximate age prediction, where tolerance of one age class is allowed. Interestingly, the inclusion of the face images modality in the PARAFAC2 model enhances the system’s performance. More specifically, the approximate age prediction performance of the speech based PARAFAC2+SVM system is improved when the face image information is included in the PARAFAC2 model (4th and 5th lines of third column of Table 5.4). Furthermore, when PARAFAC2 is trained on both speech and face image features, the SVMs applied solely on speech ranking vectors perform better (0.3040) from the SVMs trained on both speech and image ranking vectors (0.2247).

<table>
<thead>
<tr>
<th>Age prediction $F_1$ measure results</th>
<th>Age prediction</th>
<th>Approximate Age prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARAFAC2-Speech, No SVM</td>
<td>0.2028</td>
<td>0.3188</td>
</tr>
<tr>
<td>PARAFAC2-Speech, SVM-Speech</td>
<td>0.2996</td>
<td>0.5198</td>
</tr>
<tr>
<td>PARAFAC2-Speech+Image, SVM-Speech</td>
<td>0.3040</td>
<td>0.5683</td>
</tr>
<tr>
<td>PARAFAC2-Speech+Image, SVM-Speech+Image</td>
<td>0.2247</td>
<td>0.5374</td>
</tr>
</tbody>
</table>

Table 5.4 $F_1$ measure results for age prediction derived from the unimodal system presented in Section 4, a PARAFAC2+SVM system where PARAFAC2 is trained on speech feature vectors (3 slices) and SVM on speech ranking vectors, a PARAFAC2+SVM system where PARAFAC2 is trained on speech and face image feature vectors (4 slices) and SVM on speech ranking vectors and a PARAFAC2+SVM system where PARAFAC2 is trained on speech and face image feature vectors (4 slices) and SVMs on speech and image ranking vectors, respectively.

Correspondingly, in Table 5.5, the experimental results of gender prediction of two proposed frameworks methods are comparatively presented. The best $F_1$ measure value (0.8370) is admitted when PARAFAC2 is based only on speech utterances and SVMs are applied on the speech ranking vectors. The results depicted in Table 5.5 are similar for the different approaches examined and no safe conclusions can be reached whether the inclusion of face images in the PARAFAC2 model or the utilization of SVMs for decision making boosts the gender prediction performance.
5.4 Conclusions

An appealing method for age interval and gender prediction has been proposed. PARAFAC2 has been employed for extracting a semantically related sketch from face image intensities and auditory temporal modulations either exclusively or jointly in a bimodal fashion. Next, ranking vectors have been derived by PARAFAC2, which have been classified by SVMs trained for age interval and gender prediction. The former is a multi-class classification problem, while the latter is a binary classification problem. The application of SVMs on ranking vectors present the

Gender prediction has been performed more accurately than age interval prediction. In both tasks, a better $F_1$ measure has been reported when speech utterances are employed using face images within the proposed method, which has been proved to be a hard task. This may be attributed to the large variability of the images’ recording conditions (i.e., lighting, background, pose or appearance) as well as the small size of the dataset. It is also worth noting that the task under study involves age prediction using narrow range intervals. Also, age prediction can be generally considered challenging since biological age may differ drastically from chronological age. Despite the aforementioned degrees of difficulty, humans have been proven more efficient in predicting the age interval when they are looking to the

<table>
<thead>
<tr>
<th>Gender prediction $F_1$ measure results</th>
<th>Gender prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARAFAC2-Speech, No SVM</td>
<td>0.8159</td>
</tr>
<tr>
<td>PARAFAC2-Speech, SVM-Speech</td>
<td>0.8370</td>
</tr>
<tr>
<td>PARAFAC2-Speech+Image, SVM-Speech</td>
<td>0.8018</td>
</tr>
<tr>
<td>PARAFAC2-Speech+Image, SVM-Speech+Image</td>
<td>0.8282</td>
</tr>
</tbody>
</table>

Table 5.5 $F_1$ measure results for gender prediction derived from the unimodal system presented in Section 4, a PARAFAC2+SVM system where PARAFAC2 is trained on speech feature vectors (3 slices) and SVM on speech ranking vectors, a PARAFAC2+SVM system where PARAFAC2 is trained on speech and face image feature vectors (4 slices) and SVM on speech ranking vectors and a PARAFAC2+SVM system where PARAFAC2 is trained on speech and face image feature vectors(4 slices) and SVMs on speech and face image ranking vectors, respectively.
face images than listening to speech utterances.
PARAFAC2 is a multi-way decomposition method that is able to perform semantically oriented dimensionality reduction and extract latent features from multi-way data. In this thesis, a novel framework for gender and age interval prediction based on PARAFAC2 has been introduced. The prediction of biometric characteristics, such as gender and age, are important aspects for a typical biometric verification system. The contribution of this thesis is the exploitation of the powerful decomposition properties of PARAFAC2 in soft biometrics. To this scope, two biometric systems are proposed: a unimodal and a bimodal. The unimodal biometric system predicts gender and age interval based on speech utterances, while the bimodal relies on both speech utterances and face images in order to make predictions. Moreover, the bimodal biometric system incorporates SVMs applied on the age interval and gender prediction ranking vectors derived via PARAFAC2 into the decision making process.

The experimental evaluation of the proposed unimodal and bimodal frameworks indicates the strength of PARAFAC2 to yield solid predictions for the biometric characteristics under investigation. Both systems achieve better results for gender prediction rather than age interval estimation. Apparently, the age prediction problem seems to be quite more
challenging for the proposed frameworks. This can be justified since, even for humans, determining someone’s gender is generally considered easier than determining someone’s age.

It is worth noting that a straightforward comparison between the performance of the unimodal and the bimodal system is not quite fair, since different frameworks are utilized and the systems are evaluated on somewhat different datasets, due to the lack of contemporary face images for all the recordings of the TCDSA dataset. Nevertheless, while taking the aforementioned considerations into account, a few interesting observations can be noted. The utilization of SVMs in the decision making process seems to improve the system’s performance on age prediction. Moreover, the fact that the bimodal system’s performance in age prediction improved with the inclusion of the face images modality in the PARAFAC2 model is promising. Furthermore, the proposed framework that utilizes both PARAFAC2 and SVMs seems to perform better based on speech rather than based on face images. Interestingly, the opposite is the case for humans, as indicated by the questionnaire results, i.e., humans perform better in age prediction based on face images rather than based on speech recordings. The gender prediction results are quite similar for both proposed systems and demonstrate a solid gender prediction approach.

The experimental results presented in this thesis evince the competency of PARAFAC2 to capture hidden relationships among the speech recordings and face images. An interesting direction for future work involves the investigation of different ways of exploiting the ranking vectors provided via PARAFAC2 in order to make predictions. Moreover, the intriguing possibility of applying PARAFAC2 in a regression framework for the prediction of age can be explored.
Bibliography


