Selected Topics in Audio-based Recommendation of TV Content

Shepstone, Sven Ewan

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Selected Topics in Audio-based Recommendation of TV Content

Ph.D. Dissertation

Sven Ewan Shepstone

Dissertation submitted March 9, 2015

Selected Topics in Audio-based Recommendation of TV Content

Thesis submitted:	March 9, 2015
PhD Supervisor:	Assoc. Prof. Zheng-Hua Tan Aalborg University
PhD Co-supervisor:	Prof. Søren Holdt Jensen Aalborg University
PhD Company Supervisor:	Thomas Fiil Bang & Olufsen A/S
PhD Committee:	Assoc. Prof. Lars Bo Larsen, Aalborg University (committee chairman) Dr. Najim Dehak, Massachusetts Institute of Technology Prof. Lars Kai Hansen, Technical University of Denmark
PhD Series:	Department of Electronic Systems, Aalborg University

This work is supported by Bang and Olufsen A/S, Denmark and the Danish Ministry of Science, Innovation and Higher Education under Grant no 12-122802.

ISBN: 978-87-7152-068-2

March 2015

Department of Electronic Systems
Aalborg
Fredrik Bajers Vej 7
DK 9220
Denmark

This manuscript was written in LaTeX 2_ε.

Abstract

The central theme of this thesis is the recommendation of TV programs or enhancing the Electronic Program Guide (EPG) using audio-derived profiles constructed using text-independent speaker recognition techniques. For recommendation purposes, the focus is primarily on utilizing age, gender, and emotion classification. This way of implicitly deriving user parameters from audio speech is in contrast to using ratings, usage patterns, or other explicitly provided user data, which are common techniques for recommender systems. In addition to using classification to enhance recommendation, the thesis also focuses on the core technology of speaker recognition, and proposes a novel way to address source variation.

The first part of the PhD thesis focused on using state-of-the-art age and gender detection of groups of viewers to recommend sequences of short advertisement TV clips to them. A classifier using acoustic as well as prosodic features was used to classify seven age and gender classes. The major finding here was that advertisements presented using audio-derived demographics for a group of viewers in front of the TV received higher ratings than randomly selected advertisements.

The second part of the thesis incorporated discrete user emotions in a framework for recommending content using a browsing based approach. An i-vector based system was used to detect twelve discrete emotion states, which, over time, were condensed to a single mood state. The detected mood was mapped to the closest matching item in the valence-arousal space. If the initially proposed item did not match a user's personal taste for a given mood, critiquing was used to give users the opportunity to find an alternative item, if desired. Once the user had located the final item, the offset (if any) from initial item to final item was measured, and stored. This offset was termed *affective offset* and was used in future recommendation rounds to suggest a more suitable recommendation for the subject. The major finding here was that there was an increase in user satisfaction, where the average ratings for items went up when recommending using the affective offset approach. Furthermore, when affective offset was applied, there was a decrease in the number of iterations needed to find a more suitable item.

In the third part of this thesis, a novel framework was introduced that combined the strength of machine-based and human emotion classification. Based on a person's measured emotion granularity, a personalized adapted audio emotion

classifier was proposed. The system was evaluated from a similarity of emotions perspective, which subjects rated themselves. The most important finding was that for high-granularity subjects, granularity-adapted emotion classification was able to beat a 12-class baseline classifier, by improving the potential similarity for these subjects.

In the fourth and final phase of the thesis, the thesis looked at minimizing source variation, from the perspective of speaker recognition, and in particular the focus was on total variability modeling (i-vectors). The focus therefore shifted from recommendation to proposing enhancements for the core technology of speaker recognition. In a real-life scenario, there is likely to be a mismatch between the test speech and the speech of the target model. In standard total variability modeling, the prior for training the total variability matrix and extracting i-vectors is non-informative, since test and target data are assumed to come from the same homogeneous source. For heterogeneous datasets, using an informative prior instead is applicable, and it was shown here that this can lead to favorable results. The results were evaluated using the NIST 2008 and NIST 2010 Speaker Recognition Evaluation (SRE) dataset, and the proposed method using informative priors was compared to four baselines. The findings showed that whether informative priors are limited to extracting i-vectors, or used as well in training the total variability matrix, the proposals beat the baselines in a substantial number of common conditions, for both SRE'08 and SRE'10.

Resumé

Afhandlingens centrale tema omhandler anbefaling af TV-programmer og forbedringer af den elektroniske program guide (EPG). Dette gøres ved hjælp af audio-baserede profiler, som opbygges ved brug af tekst-uafhængige talergenkendelses-metoder. Hvad angår selve anbefalingen af programmer, er det primære fokus at anvende genkendelse af seerens alder, køn og følelser. Denne implicitte måde, hvorpå man anvender parametre udledt fra talesignaler, er en ændring i forhold til de konventionelle metoder anvendt af anbefalingssystemer, der er baseret på ratings, brugsmønstre eller andre eksplicit opnåede brugerdata. Ud over at forbedre kvaliteten af selve anbefalingerne omhandler afhandlingen også den grundlæggende teknologi i talergenkendelse samt en nyskabende algoritme til at håndtere kilde-variation.

Den første del af PhD-afhandlingen omhandler anvendelse af seneste udvikling inden for genkendelse af alder og køn i grupper af seere for at anbefale sekvenser af korte reklamefilm til dem. Både akustiske og prosodiske parametre bliver anvendt til opdeling i syv klasser af alder og køn. Hovedresultatet var at reklamefilm udvalgt på baggrund af en gruppe af seeres audio-udledte demografi, førte til højere ratings end vilkårligt udvalgte reklamefilm.

Afhandlingens anden del beskriver integrering af brugeres diskrete emotionelle reaktioner i et navigationsbaseret anbefalingsframework. Et i-vektorbaseret system anvendes til at detektere tolv følelsesmæssige tilstande som over tid samles til en enkelt humørtilstand. Det registrerede humør afbildes som det nærmeste punkt i valens-ophidselsesrummet. Hvis det først foreslåede program ikke falder i brugerens personlige smag for et givet humør, anvendes 'critiquing'-metoden til at finde et alternativ. Når brugeren finder det endelige program, måles og gemmes afstanden fra første til sidste program. Denne afstand kaldes for *affektiv afstand* og bruges i fremtidige anbefalinger til at komme med bedre anbefalinger til brugeren. Hovedresultatet var en øget brugertilfredshed, hvor gennemsnitsratings for programmer blev forøget når anbefalinger var baseret på affektiv-afstandsmetoden. Ydermere medførte affektiv-afstandsmetoden et fald i det nødvendige antal af iterationer for at finde et mere passende program.

I afhandlingens tredje del introduceres et nyskabende framework som kombinerer styrkerne ved maskinebaseret og menneskelig emotionel klassificering. Ba-

seret på en persons målte følelsesselektivitet foreslås en adapteret audioemotionel klassificering. Systemet blev evalueret ud fra lighed mellem følelser, som forsøgspersonerne selv vurderede. Hovedresultatet var en forøget potentiel lighed, for forsøgspersoner med høj følelsesselektivitet, når en selektivitetsadapteret følelsesklassificering blev sammenlignet med en tolv-niveaus reference klassificering.

I den fjerde og sidste fase af afhandlingen bliver der kigget på at minimere kildevariationen i forhold til talergenkendelse, med særligt fokus på total-variationsmodellering (i-vektorer). Fokus ændres dermed til at foreslå forbedringer til den grundlæggende teknologi i talergenkendelse. I et virkeligt scenarie vil der ofte være forskel på den tale der undersøges, og den der sammenliges med. I almindelig total-variationsmodellering vil prior-fordelingen, til at træne total-variationsmatricen og udtrække i-vektorer, være ikke-informativ eftersom test-dataen antages at komme fra den samme homogene kilde som den data der sammenlignes med. For et heterogent dataset er det muligt at anvende en informativ prior-fordeling, og det blev vist at dette giver en forbedring. Resultaterne blev evalueret ved anvendelse af NIST 2008 og NIST 2010 Speaker Recognition Evaluation (SRE) datasættet, hvor den foreslåede metode der anvender informativ prior-fordeling sammenlignes med fire referencer. Resultaterne viste at hvad end den informative prior-fordeling var begrænset til udtræk af i-vektorer eller også anvendes i træningen af total-variationsmatricen, gav de foreslåede algoritmer væsentligt bedre resultater end referencerne i forhold til en række fælles konditioner, for både SRE'08 og SRE'10.

Contents

Abbreviations

ASR	Automatic Speech Recognition
CC	Common Condition
DCF	Detection Cost Function
DET	Detection Error Tradeoff
DNN	Deep Neural Network
EER	Equal Error Rate
EM	Expectation Maximization
EPG	Electronic Program Guide
GMM	Gaussian Mixture Model
GVC	Group Viewer Configuration
GEMEP	Geneva Multimodal Emotion Portrayals
IAPS	International Standard Picture Database
ISV	Intersession Variability
JFA	Joint Factor Analysis
LDA	Linear Discriminant Analysis
LDC	Linear Discriminant Classifier
LLD	Low Level Descriptor
MAP	Maximum-A-Posterior
MDS	Multidimensional Scaling
MFCC	Mel Frequency Cepstral Coefficients
MLP	Multilayer Perceptron
NIST	National Institute of Standards and Technology
PCA	Principal Component Analysis
PLDA	Probabilistic Linear Discriminant Analysis
PLP	Perceptual Linear Prediction
SAM	Self Assessment Manikin
SRE	Speaker Recognition Evaluation
SNAW	Source Normalized and Weighted
SVM	Support Vector Machine
UBM	Universal Background Model
VA	Valence Arousal
WCCN	Within-Class Cosine Normalization

List of Publications

Thesis Title:	Selected Topics in Audio-based Recommendation of TV content
Ph.D. Student:	Sven Ewan Shepstone
University Supervisors:	Assoc. Prof. Zheng-Hua Tan, Aalborg University Prof. Søren Holdt Jensen, Aalborg University
Company Supervisor:	Thomas Fiil, Bang & Olufsen A/S

This thesis is based on the following publications:

[A] S.E. Shepstone, Z-H. Tan, S. H. Jensen, "Demographic Recommendation by means of Group Profile Elicitation Using Speaker Age and Gender Recognition," *Proc. INTERSPEECH*, pp. 2027–2031, 2013.

[B] S.E. Shepstone, Z-H. Tan, S.H. Jensen, "Audio-based Age and Gender Identification to Enhance the Recommendation of TV Content," *IEEE Trans. on Consumer Electronics*, vol. 50, no. 3, pp. 721–729, 2013.

[C] S.E. Shepstone, Z-H. Tan, S.H. Jensen, "Using Audio-derived Affective Offset to Enhance TV Recommendation," *IEEE Trans. on Multimedia*, vol. 16, no. 7, pp. 1999–2010, 2014.

[D] S.E. Shepstone, Z-H. Tan, S.H. Jensen, "Audio-based Granularity-adapted Emotion Classification," Submitted to *IEEE Trans. on Affective Computing*.

[E] S.E. Shepstone, K. A. Lee, H. Li, Z-H. Tan, S.H. Jensen, "Source-specific Informative Prior for i-Vector Extraction," Accepted for *Int. Conf. Acoust., Speech, Signal Process.*, April 2015.

[F] S.E. Shepstone, K. A. Lee, H. Li, Z-H. Tan, S.H. Jensen, "Total Variability Modeling Using Source-specific Priors," Submitted to *IEEE Trans. Audio, Speech and Lang. Process.*.

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted or published scientific papers which are listed above. Parts of the papers are used directly or indirectly in the extended

summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty. The thesis is not in its present form acceptable for open publication but only in limited and closed circulation as copyright may not be ensured.

To Jeanine, Daniel and Linus.

Preface

This thesis is submitted to the Doctoral School of Engineering and Science at Aalborg University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. This thesis falls within the framework of an industrial PhD project, which is a collaboration between Aalborg University and Bang and Olufsen A/S, where the student was employed throughout the project's duration. The work was carried out in the period from June 1, 2012 to March 9, 2015, at Bang and Olufsen as well as at the Department of Electronic Systems at Aalborg University.

The thesis is concerned with recommending TV programs or enhancing the Electronic Program Guide (EPG) using profiles built from audio-derived parameters from speech. The profile attributes are obtained using classification based on text-independent speaker recognition techniques. In addition to using classification to enhance recommendation, the thesis also focuses on aspects relating to enhancements of the core technology of speaker recognition itself. In the first part we present an overview, the current state-of-the-art, and the project's major contributions. In the second part, the papers relevant to this thesis are enclosed. At the time of submission of this thesis, four papers were accepted for publication, and two papers were undergoing peer-review. The papers have been ordered chronologically.

I would like to express my gratitude and appreciation to Bang and Olufsen as well as the Danish Ministry of Science, Education and Higher Education for funding this research. There are also a number of individual people that deserve a big thank you. Firstly I would like to thank my main supervisor Assoc. Prof. Zheng-Hua Tan for all his help and the invaluable time he invested, for providing top quality supervision, for reviewing the various papers of this thesis throughout the course of the PhD, and last, but not least, for making my time at Aalborg University a thoroughly enjoyable experience. I also want to thank him for setting up my four-month visit to the Human Language Technology department at the Institute for Infocomm Research at A'STAR in Singapore for the summer of 2014. I would like to thank my co-supervisor Prof. Søren Holdt Jensen for his supervision, for reviewing all the papers and for his valuable feedback and suggestions. With respect to my overseas visit, I would like to sincerely thank both Dr. Kong Aik Lee and Dr. Haizhou Li for hosting my visit in Singapore and for dedicating their precious

time to me. I want to thank Thomas Fiil at Bang and Olufsen for his support throughout the project, and for his advice. I also want to thank Assoc. Prof. Lars Bo Larsen at Aalborg University for fruitful discussions on conducting user studies and Ditte Hvas Mortensen for her assistance in a number of the psychological matters relating to the thesis. Finally I would like to thank all the members of my department at Bang and Olufsen as well as Aalborg University for supporting me throughout the project. I want to thank Jan Thunbo Brander, Søren Bech and Søren Borup Jensen for their support and many good discussions. Thank you to Jesper Kær Nielsen for his assistance with typesetting this manuscript in LaTeX 2_ε, and for making such a great template available. Thank you also to Nicolai Bæk Thomsen for good discussions on factor analysis. Thank you to Martin Bo Møller for helping to write a good-quality Danish abstract. Last, but certainly not least, I want to thank my wonderful family Jeanine, Daniel and Linus for supporting me tremendously throughout the past three years - words can not express how grateful I am to you for the sacrifices you've made on my behalf.

Sven Ewan Shepstone
Bang and Olufsen / Aalborg University, March 9, 2015

Part I

Introduction

Audio-based Recommendation of TV Content

1 Introduction

1.1 Problem Statement

The volume of media content has increased dramatically in recent years, and therefore a large body of research has looked into ways of assisting users on what to watch. In the context of broadcast TV, the Electronic Program Guide (EPG), now an integral part of most of today's set-top boxes and television sets, provides a limited amount of assistance, but even with the short time span covered by the EPG, there can still be an overwhelming number of programs to choose from. Recommender systems [1–3], which have been very successful for on-demand movie items such as those found on Netflix, tend to be overlooked when it comes to the broadcast domain.

In order to recommend something personal, a user profile is needed. A profile is simply a generalization of information about the user, and captures the needs and desires of the customer from the perspective of recommendation [4]. In collaborative filtering, this user profile information takes the form of explicitly provided ratings. In content-based filtering, the user profile could contain keywords obtained from content items that the user rated highly in the past, or as obtained by means of a registration questionnaire [5]. In the context of the EPG framework, there have been works that have proposed collaborative [6] as well as content-based techniques [7]. One of the problems seen with explicit ratings is that users do not always take the time to rate content, and there has therefore been longstanding interest in how best to minimize this rating effort [8]. It is also possible to build a user profile implicitly, such as through usage patterns [6]. When multiple users share a single device such as a television, not only must individual interests be catered for, but also the group as a whole [9].

Instead of explicitly provided data, or implicitly provided usage patterns, this work proposes using audio analysis techniques for recommendation. More specifically, the focus in this interdisciplinary field is on how paralinguistic phenomena extracted from a person's speech signal can be used to enhance the recommendation of TV content. Paralinguistic phenomena refer to the non-linguistic (*alongside linguistics*, hence *para*) cues of a person's speech [10], from which a wealth of information about the user can be extracted. This information can then form the basis for recommendation. Our proposals for recommendation are within the framework of broadcast TV, where classical techniques for recommendation might meet certain limitations. For example, other peoples' ratings have limited use in the context of broadcast TV for content that will only ever be shown once, and when showing content to groups of viewers, it cannot always be assumed that those watching will identify themselves upfront.

1.2 Main Hypotheses

This thesis concerns itself with two main hypotheses. Firstly we test the hypothesis that speaker classification techniques using paralinguistic information from the speech signal of TV viewers can be used to recommend content items to them. This is a rather broad hypothesis and in Section 6 we shall break this more general hypothesis down into a series of more specific sub-hypotheses, each connected directly with the specific proposal in mind. We ask the reader to bear with us until then. Secondly, the very application of extracting audio parameters from TV viewers can result in mismatches between the speech used to train speaker recognition target models and the speech from the test environment. Such a target model could for example represent the identity of a speaker, or a speaker class, such as a specified age group. We cannot assume that the type of speech used for training these target models will be the same as that under actual use. This very fact calls for a robust speaker recognition framework to be considered. When the speech for the target and test models comes from multiple sources, this is known as source variation. In this light we test the hypothesis that domain knowledge obtained about the type of speech in question can be used to minimize this source variation. This added robustness would most certainly be an advantage when carrying out speaker classification.

1.3 Outline

The rest of this introductory chapter and extended summary for the thesis is structured as follows: Since the technology of speaker recognition is central to this thesis, we begin in Section 2 by giving an introduction to the state-of-art in the field, with particular emphasis on recent trends and research directions, with no particular application in mind. The following section focuses on speaker classification, where after introducing basic emotion theory, we present the state-of-the-art

in identifying emotions, age and gender. Section 4 introduces recommender systems and we examine their usage in some interesting frameworks. By structuring these sections in the manner we have done, we hope to fulfill our intention of guiding the reader from the abstract to the applied, from the perspective of the central ideas of this thesis. In Section 5, we present a framework for how audio-based recommendation of TV content might be accomplished. As this research is fairly controversial, a short sub-section on privacy and ethical issues is included. The next section presents an overview of the major contributions, along with a sub-hypothesis for each. The final section gives concluding remarks, as well as a number of proposals for possible future work in the area.

2 State-of-the-art in Robust Speaker Recognition

Speaker Recognition is the process of recognizing people from their voice, where the physical voice characteristics as well as each person's manner of speaking plays a role [11]. We give a brief overview of the major advances in speaker recognition over the past 15 years with particular emphasis on total variability modeling, which is the most recent development in the state of the art.

2.1 UBM-GMM Modeling

In the UBM-GMM approach a Gaussian Mixture Model (GMM) known as a Universal Background Model (UBM) is trained, usually using a large amount of neutral speech [11, 12]. The Expectation-Maximization (EM) algorithm [13] is used to determine the weights, means and covariances. For each iteration of the EM algorithm, the likelihood of a lower bound function, parameterized by the model parameters, increases. The algorithm continues until convergence. It is assumed that this UBM is multivariate with F dimensions and C mixtures. For each target speaker or class that is to be modeled, a separate adapted GMM is trained. Each GMM can be obtained by adapting the weights, means and covariances of the UBM and the class's data using the maximum-a-posterior approach (MAP). However, in text-independent speaker recognition, typically only the means are adapted. A relevance factor controls the rate at which the data is shifted away from the UBM. To evaluate whether a speech utterance belongs to a target speaker or not, or to determine which class the utterance belongs to, one calculates a likelihood ratio statistic which is the difference between the log likelihoods of the adapted GMM and UBM.

2.2 Modeling with Supervectors

It is possible to concatenate the means of an adapted GMM to form a high dimensional $C \times F$ supervector. In other words, each utterance can now be represented by a fixed-length, high-dimensional supervector. Once in the supervector space, a linear discriminant classifier such as a support vector machine (SVM) can be used [14, 15].

2.3 JFA Modeling

In Joint-Factor Analysis (JFA) [16, 17], factor-analysis techniques are used to represent the GMM-extracted supervector $\mathbf{m}$ of a speaker utterance in the following additive way:

$$\mathbf{m} = \mathbf{m_0} + \mathbf{Vy} + \mathbf{Ux} + \mathbf{Dz} \tag{1}$$

where $\mathbf{m_0}$ is the speaker-independent supervector (this supervector is obtained by stacking the F-dimensional mean vectors of the UBM [11]). The speaker supervector component is $\mathbf{Vy} + \mathbf{Dz}$. Here $\mathbf{V}$ and $\mathbf{D}$ together define the speaker sub-space. $\mathbf{V}$ is an eigenvoice matrix and its columns are the eigenvoices. $\mathbf{D}$ is a diagonal matrix and models the residual speaker variability not modeled by $\mathbf{V}$. The channel supervector component is $\mathbf{Ux}$, where the matrix $\mathbf{U}$ defines the session-subspace. These matrices are also trained beforehand using a variant of the Expectation Maximization algorithm [13] and the process is outlined in [18]. A common practice is to first train $\mathbf{V}$, followed then by $\mathbf{U}$ and $\mathbf{D}$. The speaker factors $\mathbf{y}$ and $\mathbf{x}$ are hidden random variables each assumed to have a standard normal distribution $\mathcal{N}(0, \mathbf{I})$, and which define the location of each utterance in the respective subspace. The unwanted variability can then be removed when scoring a test-utterance against a target model.

2.4 Total Variability Modeling

Total variability modeling has been the state of the art in speaker recognition in the last few years. In total variability modeling [19], variable-length speech utterances are mapped from the high-dimensional supervector space to fixed low-dimensional i-vectors. This low-dimensionality and fixed length, along with the fact that the actual i-vectors are generated via unsupervised learning [20], means that tasks relating to classification, identification and verification of speakers are greatly simplified.

Assuming an F-dimensional feature space, and a GMM with C components, the total variability model assumes that a GMM-extracted supervector $\mathbf{m}$ for the speech utterance can be represented as:

$$\mathbf{m} = \mathbf{m_0} + \mathbf{Tw} \tag{2}$$

where $\mathbf{m_0}$ is the speaker-independent supervector (this supervector is obtained by stacking the F-dimensional mean vectors of the Universal Background Model (UBM) [11]), $\mathbf{T}$ is a matrix of low rank and $\mathbf{w}$ is a hidden random variable assumed to have a standard normal distribution $\mathcal{N}(0, \mathbf{I})$. The supervector $\mathbf{m}$ is assumed to be normally distributed with mean $\mathbf{m_0}$ and covariance $\mathbf{T}\mathbf{T}^{\mathrm{T}}$. The i-vector is then just the MAP point estimate of this hidden variable. Similarly to JFA, the matrix $\mathbf{T}$ is again trained using the EM algorithm approach for eigenvoice matrices [18], except that utterances from the same speaker are now treated as if coming from individual speakers. By doing this, the $\mathbf{T}$ matrix not only captures the speaker variability, but also the channel variability.

Given a set of observations representing the feature sequence of an utterance $\{\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, ...\mathbf{o}_T\}$, the following zero-order Baum-Welch statistics are defined:

$$N(c) = \sum_t \gamma_t(c) \tag{3}$$

where t is the frame index and $\gamma_t(c)$ is the occupancy of the t^{th} frame to the c^{th} Gaussian. The first-order statistics, which have also been centered, are defined as:

$$\widetilde{\mathbf{F}}(c) = \sum_t \gamma_t(c)(\mathbf{o}_t - \mathbf{m_0}(c)) \tag{4}$$

Furthermore, let $\mathbf{N}$ represent the $CF \times CF$ diagonal matrix with blocks given by $N(c) \times \mathrm{I}$ and let $\widetilde{\mathbf{F}}$ represent the $CF \times 1$ supervector obtained by stacking the $\widetilde{\mathbf{F}}(c)$ for all C.

In the E-step, we compute the posterior distribution for the hidden variables given the set of observations $\mathbf{O}$, and a non-informative prior, i.e. a standard normal prior $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$:

$$p(\mathbf{w}|O) = \mathcal{N}(\mathbf{L}^{-1} \cdot \mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}},\ \mathbf{L}^{-1}) \tag{5}$$

with mean vector $\phi = \mathbf{L}^{-1} \cdot \mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}}$ and precision matrix $\mathbf{L} = (\mathbf{I} + \mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{T})$.

In the M-step, assuming I training utterances, the current value for $\mathbf{T}$ is updated by solving a set of simultaneous equations:

$$\sum_i \mathbf{N}_i\mathbf{T}.[\phi_i{\phi_i}^{\mathrm{T}} + \mathbf{L}_i] = \sum_i \widetilde{\mathbf{F}}_i\phi_i \tag{6}$$

The process for extracting i-vectors is identical to the E-step above, where the mean ϕ is taken to be the i-vector. Therefore extracting an i-vector can be seen as a MAP adaptation of $\mathbf{w}$ in the total variability space. Once in the i-vector space, a number of methods can be used for carrying out scoring, including cosine-distance scoring and Probabilistic Linear Discriminant Analysis (PLDA).

The i-vectors representing speech utterances are characterized by both speaker and channel variability. As we saw for JFA [16, 17], separate factors model the speaker and channel variability, which is dealt with directly in the high-dimensional

supervector space. In total variability modeling, however, the extracted i-vectors represent the total variability (hence the name *front-end* factor analysis), and it is in this i-vector space where the unwanted channel variability is to be removed. If this channel variability is not dealt with it will negatively affect the performance of the system. Linear Discriminant Analysis (LDA) [21], WCCN [22] and PLDA [23–26] are all well-known techniques employed to reduce channel variability. PLDA has been shown to exhibit superior performance for i-vectors and is now the standard back-end for reducing variability and computing verification scores in the i-vector space.

In the early days of speaker recognition, the speech samples used for evaluation were predominantly telephone speech. However, in more recent years more and more microphone speech as well as so-called interview speech has been recorded and included in speaker recognition evaluation corpora [27, 28]. The resulting heterogeneous datasets from these corpora will, in addition to speaker and channel variability, be characterized by source variability. This is also exactly the type of issue that would occur in a real-world scenario. If this source variability is not addressed, it will impact the performance of the speaker recognition system.

One area where the issue of source variability is particularly noticeable is in the application of LDA directly in the i-vector space. Applying such an LDA transform has the unwanted effect of optimizing source variability as speaker variability. This is particularly noticeable in the inter-speaker scatter matrix $\mathbf{S}_B$. Another issue encountered with LDA is that a lack of utterances from all sources for each speaker leads to poor estimation of the within-speaker scatter matrix $\mathbf{S}_W$. These issues were addressed in [29, 30], by modifying the standard LDA algorithm with a source-normalized and weighted approach (SNAW), that leads to better estimation of both the $\mathbf{S}_B$ and $\mathbf{S}_W$ scatter matrices.

In addition to source mismatch, an issue often encountered is that the amount of the telephone data available for training models is often substantially more than data for other sources. Another method that has been proposed to deal with source variability is to train a total variability matrix using only telephone speech, and then a lower-dimensional supplementary matrix using only microphone speech [21]. In the training of the supplementary matrix, the already-trained telephone matrix is included as an additive factor in representing the speaker and channel dependent supervector $\mathbf{m}$. Before extracting i-vectors, these two matrices are concatenated to form a cascade system. A problem with this approach is that the space spanned by the i-vectors are different - i-vectors from telephone-acquired speech reside in a smaller space than the i-vectors from microphone-acquired speech. This issue was addressed by using PLDA to project the two spaces to a single space [31].

I-vectors have shown great improvements in performance over existing techniques such as GMM-SVM and JFA. Very recently, it was discovered that additional performance can be obtained by replacing the UBM with a deep neural network (DNN) with no alteration to the i-vector modeling itself [32]. The use of DNNs was inspired by their performance success for speech recognition as well as the

fact that speech content is largely ignored for speaker recognition. Senones (tied triphone states) obtained from a decision tree, are used as classes in a DNN trained for Automatic Speech Recognition (ASR). By allowing these senones to replace the Gaussian posterior probabilities from the UBM, the Baum-Welch statistics can then be determined from these senone posteriors instead. This structure allows more meaning to be attributed to the posterior probabilities as speakers are now able to be compared over the same phonetic units. It also allows decoupling (and a more optimum selection) of the features used to generate the probabilities, and those used to extract the i-vectors. In [33] it was found that performance depends to a large extent on what data is used to train the DNN. By extending the framework to microphone speech, it was demonstrated that including telephone speech into the training of the DNN led to significant performance improvements for microphone test data over a series of baselines, indicating that the added telephone data adds better generalization ability.

3 Speaker Classification

Speaker classification is defined as the process of assigning a single speech utterance to a given class. From a hierarchical perspective, speaker recognition may be defined as a sub-field of speaker classification [34]. In speaker classification, the classes are described by paralinguistic phenomena, which concern all aspects of speech that cannot be described using linguistic or phonetic cues [10]. Such phenomena can be long-term traits, such as age and gender, or short term states, such as a person's affective state, and occur at both acoustic and linguistic levels.

The recent surge of interest in Human-Machine communication has resulted in a strong motivation for extracting paralinguistic information since there are many applications that can benefit from it [10]. These include age and gender detection to enhance speech recognition (by employing the appropriate model), emotion detection for call centers (for example, passing an angry customer's call to an appropriate person), analyzing affective states to improve human-robot interaction, conveying paralinguistic cues such as emotion to autistic children, detecting intoxication levels for law enforcement and so forth. Paralinguistics are also used in the field of multimedia retrieval, to find highlights in sports games. Our interest is primarily in the entertainment domain, where the major focus is on how peoples' emotions, age and gender can be detected from speech. We begin by giving a short overview of emotion modeling, within the more general context of its application to digital systems as a whole, before moving on to the state-of-the-art in automatic detection of emotions from speech.

3.1 Emotion Modeling Applied to Human-Computer Interaction

There is no one unified model for handling emotions [35]. One way to model emotions is to treat them as categories, and there is evidence to suggest that most people possess six basic emotions [36]. These emotion categories are often referred to in the literature as Ekman's Big Six and are happiness, sadness, anger, fear, surprise and disgust. Modeling of emotions as categories is especially popular in the field of speech processing, where emotion states are given labels. This is partially due to historical reasons and was motivated by the type of corpora that initially became available [37].

In the Circumplex Model of Affect, discrete emotional states are expressed as a circular model in two-dimensional space occupied by valence and arousal [38]. In this model, emotions are ordered around the circumference of a circle, with similar emotions adjacent to one another and bipolar emotions on opposite sides of the circle. Empirical studies have shown that emotions can be characterized by their location around the circle in degrees [38, 39].

In the dimensional approach for representing emotions, valence refers to the pleasantness of the emotion experienced, whereas arousal, also called activation, refers to the degree of arousal experienced (sometimes this is referred to as intensity). It has been suggested that emotional feelings are primarily characterized by valence and secondarily by activation [39]. There is also evidence to suggest the presence of another dimension, namely dominance, also called potency [40]. Dominance refers to the degree of control of the emotion experienced, for example, anger (more control) and fear (less control). However, since the control dimension only plays a small role in characterizing emotional states [41], for simplicity, many modeling approaches ignore the dominance dimension altogether.

According to the Circumplex Model of affect, each emotion can be characterized by a certain degree of valence and arousal, and it is possible to portray these differences in geometric space. In the ideal case, an individual will weight the valence and arousal dimensions equally. In particular, this would mean that for two emotions that have identical valence and differ only in arousal, the individual would just as easily be able to tell apart two emotions that differ valence-wise, with identical arousal values. However, this is rarely the case, and most people differ to the extent in which they weight either the valence or arousal dimension. A person who places a lot of emphasis on valence is said to have a high valence focus, and when a person places emphasis on arousal, they are said to have a high arousal focus [42]. The ability to distinguish emotions from one another is known as emotion granularity [43]. By carrying out a series of experiments that record people's self-reports of emotion, it is possible to compute a person's granularity. The differing abilities of individuals with regard to their granularity makes sense since it fits well with the inherent taxonomy of emotions: At the highest level are main categories such as positive and negative, followed by more general discrete

labels, such as Ekman's Big 6 emotions at the next level, finally followed by sub-categories at the lowest level [37].

When it comes to selecting content, mood management theory suggests that people generally follow their hedonistic ideas, and will select content that keeps them in a good mood [44]. However, not everyone agrees on what content is applicable for a given mood setting [45, 46]. For example, for two people who are in a bad mood, one might request drama to remain in that bad mood whereas the other might request comedy to repair their bad mood state.

To make emotions useful in modern day digital systems [47], the emotions of the user as well as those of the content need to be determined for a good match to be made. Popular ways to detect emotions are by using physiological measures [48], by multimodal audiovisual analysis [49], and through speech [10, 37], which is the primary focus here.

Many works have looked at extracting emotions from media content, primarily in connection with affective video content tagging, characterization and retrieval. In [41] the authors propose to use audio and video cues to derive a representation based on valence and arousal. In [50], the authors propose a dimensional approach where speech from movies is mapped directly to valence and arousal. Human-annotated coordinates for speech utterances along with the corresponding feature vectors are used to train a regression model. Test utterances are then mapped using this model directly to the valence around plane, and evaluated by comparing the performance with human-annotated data.

3.2 Automatic Detection of Emotions from Speech

Automatic emotion processing, now a mainstream research topic, is concerned with the problem of determining the correct emotional state of a person from their speech [10]. This is possible since there are a number of attributes at different levels of speech that have been found to vary with emotion, and which are more or less constant across people.

Emotion databases can be created using either acted speech [51–53] or spontaneous speech, sometimes called non-prompted speech [54, 55]. Due to privacy concerns associated with gathering spontaneous emotions, the majority of emotional databases have been constructed using acted emotions. Although spontaneously acquired data is more realistic, the data is often collected with a particular application in mind [37], which might complicate its usage in other contexts. Acoustic features have been found to be particularly beneficial for acted databases, whereas linguistic features are more suited to spontaneous databases [37]. In general, speech utterances are usually assigned to fixed emotion categories, e.g. anger and fear. Since human annotators do not always agree on the emotions when labeling, a rule-of-thumb suggestion is to use at least three annotators [10].

Features for emotion can be extracted at the frame level, or at a larger level, such as the syllable, word or even entire utterance level [10]. At the frame level,

feature vectors are usually extracted at 100 frames per second, with a window size of around 25 ms, from which a set of features or low-level descriptors (LLDs) can be extracted. Popular acoustic LLDs used for emotion modeling include Mel Frequency Cepstral Coefficients (MFCCs), including their first and second derivatives, Perceptual Linear Prediction (PLP) coefficients, pitch (F0), formants, and shimmer and jitter. These features are well known, have shown good performance for emotion modeling, and have been widely adopted. Popular linguistic LLDs include linguistic units such as words (using in key-word spotting, Bag-of-words modeling and part-of-speech (POS) tags) and N-grams, as well as other units such as pauses in speech [37]. A number of these higher-level features have been inspired by their past success for text classification [56].

At the syllable or utterance level, static feature modeling using functionals may be used to approximate the LLDs over time, a process that has proven to be very effective. Applying functionals such as mean, standard deviation, and higher-order moments has two major advantages. Firstly, functionals are able to capture events at the supra-segmental level. This is useful for prosodic features that include pitch contours, energy and duration, which complement the acoustic features [57]. Secondly, they allow for the generation of fixed length feature vectors which make modeling of the features a lot easier. Often such feature sets are obtained using brute force methods and can be very large, necessitating the need for dimensionality reduction. Here, a knowledge-based method such as feature selection is often preferred to a "blind" method such as Principal Component Analysis (PCA) to allow for extraction of more meaningful features [37].

A popular classification paradigm, especially when static features are used, is to use a linear discriminant classifier (LDC) such as an SVM, which has good generalization abilities. Non-linear classifiers such as back-propagation Artificial Neural Networks (ANNs) and deep neural networks (DNNs) [58] can also be used for static features. Dynamic classifiers such as variants of GMMs and i-vector modeling [19, 59] can also be used to carry out emotion classification.

Performance results across the board show that arousal is easier to classify than valence [58, 60], that performance is higher with acted emotions than spontaneous emotions [59], and the more prototypical the emotion labels are, the better the performance [37].

A number of works have looked at extracting emotions from speech. In [57] the authors show that it is possible to use standard speaker and language identification features and techniques based on GMMs to carry out emotion recognition. The standard UBM-GMM adaptation model is extended in a novel approach, whereby the UBM and adapted GMM supervectors are adapted during verification by making use of a pre-computed latent subspace. This is used to minimize intersession variability (ISV), for example as caused by different acoustic conditions. The features used were standard MFCCs with double-delta coefficients in one ISV system and shifted delta cepstral coefficients in another ISV system. By fusing the two sub-systems, they managed to achieve an unweighted accuracy in the 2009 Emo-

tion Challenge [61] of 41.7 % for five classes taken from the FAU Aibo Emotion corpus [54].

In [59] the authors propose a separate GMM for each emotion class from which to extract i-vectors. Each GMM is adapted from a single UBM (trained using neutral data) using class-specific speech utterances, and by adapting only the mean. For each utterance, an i-vector is extracted for each emotion class, using the emotion class's GMM and $\mathbf{T}$ matrix. These i-vectors are concatenated, meaning that for each utterance, information is encoded for every emotion class. The resulting feature vectors are modeled using an SVM classifier. Results showed superior performance of 91.1 % for four emotion class on acted data and 71.3 % on spontaneous data.

In [62] the focus was on which feature types contribute best to classification performance. Brute-force and knowledge-based features are pooled from six different sites and combined into high-dimensional feature vectors for three different feature types: 3713 acoustic LLDs, 431 linguistic LLDs and 3673 acoustically-derived functionals. Separate results for each feature type, for both the full feature set, as well as a reduced feature set, are tested on four emotion classes from the FAU Aibo Emotion corpus [54], and range between 50 % and 60 %.

3.3 Automatic Detection of Age and Gender from Speech

Age and gender detection has also received a lot of research attention in recent years. On its own, gender detection has many uses: Very commonly, it is used as a front-end to gender-dependent sub-modules, such as modules trained using male-only or female-only speech. Other uses include audio-visual search and retrieval [63], and highlighting of subtitles for the hearing impaired [64].

In the beginning, gender detection focused primarily on male and female voices. In recent years however, partially due to the increasing interest in differentiating children's voices from adults, partially due to an increase in the available training data [65, 66], and partially due to the fact that there is no statistical difference between male and female voices for children under twelve years old [67], the class *children* has come to be accepted as an additional gender class. Children's voices are characterized by amongst others: a higher fundamental frequency, higher formant frequencies, greater spectral variability and higher speaking rate. After puberty, the difference between male and female voices becomes a lot more significant.

Acoustic and prosodic features are the ones that have shown most promise with age and gender detection. The acoustic features most commonly used for age and gender detection are very similar to those used for emotion: MFCCs with their first and second derivatives and prosodic ones such as pitch (F0), formants, voice quality features such as shimmer and jitter, and speaking rate. Delta and double-delta coefficients are particularly good at detecting age, since they help to capture temporal aspects such as the different speaking rates of people, in particular children and older people [68]. These short-term low-level descriptor features

are typically either extracted at the frame level, whereas more long term (supra segmental) features are obtained by using functionals, as explained for the case of emotion detection. Also, brute-forcing of features is also sometimes employed with paralinguistics, where features that contribute best to performance are selected, and where feature sets can be very large. It is also possible to extract linguistic LLDs.

In practice, age and gender classification is carried out using a number of classification paradigms. For short-term features such as MFFCs, GMM approaches such as adapted GMM models using MAP adaptation and SVM modeling using GMM supervectors are often used [69]. For longer-term features such as prosodic features or very large feature sets, SVMs are more common [15]. Other methods include multi-layer perceptions (MLPs), and methods adopted from the ASR domain such as parallel phoneme recognizers [70].

The performance of age and gender detection is affected by a number of factors. Firstly, gender detection is generally considered a more simple task, where typically up to 30 % higher accuracy (unweighted accuracy) is reported for gender than for age [67, 71]. This reason for this difference is believed to be primarily due to pitch, which itself is highly indicative of one's gender, an easy to obtain parameter from speech [10]. The confusion often seen with age and gender classification stems from the fact that certain classes have similar voices, e.g. children and females [68]. Secondly, since the quantity of training data with labeled age and gender classes is quite low compared to corpora with hundreds or thousands of hours of training data, this naturally impacts the performance.

Many state-of-the-art systems fuse the results of individual subsystems together [67, 71]. These individual subsystems are sometimes characterized in very different ways by the training data used, the features chosen, such as acoustic or prosodic features, and the classification paradigm employed. Carrying out fusion in this manner generally improves the overall robustness and classification performance of the combined system.

In 2010, the aGender corpus [65] was provided to participants in the Interspeech Paralinguistic Challenge to enhance the development of new age and gender algorithms and to improve comparability of results [72]. Since then, a lot of recent work in the field of age and gender has focused on this corpus. The corpus contains seven age and gender classes, notably children, with 46 hours of speech and 954 speakers. Other notable corpora that have been used in research on age and gender detection include the German SpeechDat II database [73] and the VoiceClass corpus.

In [68] the authors propose separate age and gender classification systems that classify four age and three gender classes respectively. The age classifier was constructed by fusing the results from four individual sub-systems. For the first subsystem, 1582 features given to 2010 Interspeech participants [72] and taken from the aGender corpus using the OpenSMILE toolkit [74] were classified using SVM. For the second sub-system, the same aGender features were classified using an MLP

approach. The third sub-system derived PLP coefficients, delta and pitch features from aGender and two additional databases, and modeled these again using an MLP approach. The final sub-system modeled temporal structure static modulation spectrogram features [75] taken from aGender and two additional databases using the standard GMM-UBM approach. The gender classifier was constructed by fusing the results from six individual subsystems: The entire age subsystem, used for gender by reduction of seven class probabilities down to three, was reused as the first sub-system. The first three age sub-systems were replicated to form the next three sub-systems. The fifth and sixth sub-systems were implemented by replicating the third MLP-based and fourth GMM-UBM-based age sub-systems, respectively, and then adding additional data. The authors managed to achieve 51.2 % unweighted accuracy for four age classes and 83.1 % unweighted accuracy for four gender classes on the development set of the aGender dataset. In a related study [67] the authors investigated additional age-and-gender front-ends.

In [71] the authors present a seven-class state-of-the-art age and gender classifier that combines seven acoustic and prosodic sub-systems using weighted summation based fusion of the results. Each sub-system is based on a different modeling technique: standard UBM-GMM modeling, linear SVM modeling with GMM means-adapted supervectors, linear SVM modeling with Maximum Likelihood Linear Regression (MLLR) matrix supervectors, Bhattacharyya (BPP) SVM modeling with UBM Weight Posterior Probability (UWPP) supervectors, a novel sparse representation method on UWPP supervectors (where the sparse non-zero distribution of test utterances and a pre-computed over-complete dictionary are compared), a reference SVM-SMILE baseline system that uses 450 features, and finally prosody modeling using contours of pitch, formant, time-based energy and frequency domain harmonic structure energy for voiced speech segments. Results on the development set of the aGender corpus show an unweighted accuracy of 52.8 % for the age task, 81.7 % for the gender task and 50.3 % for the combined age and gender task.

4 Recommender Systems

In this section, we give a brief overview of recommender system theory. Recommender systems provide personalized recommendations to users by predicting previously unseen items that are interesting or useful to them. They now have widespread application, most notably in online marketing and entertainment. We also highlight a number of exemplary recommendation strategies that incorporate either emotions or demographics in one way or another.

4.1 Classical Recommender Systems

From a classical perspective, recommender systems are categorized into collaborative filtering systems and content filtering systems [1–3]. Originally conceived in the early 90s [76], collaborative filtering systems operate on user ratings. Users supply ratings, usually according to a psychometric scale, such as the 1-5 Likert scale [77]. By determining the correlation of users' ratings for content items with one another, it is possible for each user to find a neighborhood of other similar users (by ranking the correlations). From this neighborhood, recommendations can be computed for a user's unrated items. Note that in the collaborative approach the item content does not play a role in the prediction, meaning these types of systems do not rely on domain knowledge, thus making them popular for commercial deployment. Collaborative filtering approaches are further subdivided into memory-based algorithms, that rely on non-probabilistic real-time approaches to compute ratings, and model-based approaches, that compute offline statistical models that then predict the user ratings [78]. Three disadvantages with collaborative filtering systems are that a large user community is needed before good-quality predictions can be made, for users that have not yet rated any items, predictions cannot be made, and for users whose tastes deviate too far from the norm and for whom suitable neighbors cannot easily be found, recommendation quality might be poor.

Content-based recommenders on the other hand use similarity techniques such as cosine similarity, k-nearest-neighbor or relevance feedback [7] to match attributes from a user profile directly with attributes associated with the content [1–3]. It is of course assumed that the user and content attributes are chosen to match one another. The most common way to annotate the user profile is to extract the meta-data from content items the user rated highly in the past, implying that a history is also needed. This can be done using the item descriptors directly, for example, in the case of movies, the genre. For text documents, a more effective approach however is to use information retrieval techniques, such as variants of the vector space model [79] to identify the most informative terms that distinguish one content item (also called the *document*) from another, and then automatically learn the user profile from these terms. The way modern search engines personalize search results based on implicitly derived user profiles can be seen as content-based recommendation. It is also possible to explicitly annotate the user profile (for example, just by asking the user), implying that no history is needed, or build the profile using implicit techniques. One of the big disadvantages with content-based filtering, particularly in the case of media content, is that it is not trivial to mine semantic information from the content. Therefore many content-based systems for media rely exclusively on the meta-data that accompanies the content-item, such as the title, cast, genre and synopsis.

Some works have looked at augmenting user ratings with other user parameters. In [44] the authors present a mood-aware collaborative filtering approach. By measuring self-reports of mood for each subject, the authors determined that user

mood had an influence on movie ratings and that people select content based on their mood. Adjusted movie ratings for 16 different mood states taken from the positive and negative affect schedule PANAS-X [80] were incorporated into a collaborative filtering recommender's neighborhood calculation, which was found to perform better than a classic collaborative filtering recommender.

Recommendz is a movie recommender where instead of rating entire films, users select features from a limited set, and for each feature, rate the degree to which the feature is present in the film, as well as how positively or negatively they are perceived [81]. Collaborative recommendation is carried out by weighted averaging of three separate similarity metrics on the extended ratings data.

The AIT MOvie Recommendation Engine (AMORE) is a production system that employs an ensemble of collaborative and content-based recommenders to suggest movies to customers [82]. For the collaborative component, both price as well as the time a user takes to watch an item is taken into account, which are both then used to determine the appeal of items. Separate user-based and item-based recommendation threads work independently in parallel with one another to compute neighborhood similarities. Content-based filtering is carried out by exploiting the already available metadata of content items. The top-n recommendations for a user are then computed by taking a linear weighted combination of all three recommendation results (content-based, user-based collaborative and item-based collaborative). The results showed that AMORE was able to outperform Apache Mahout by more than 100 % in terms of recall for n=10 top items.

When it comes to recommending content at the affective level, it is a not a trivial process to construct a user profile. In [46], the authors consider how an emotion-based movie recommender system might be constructed taking into a user's movie ratings according to emotions. By detecting a person's emotion, other like-minded peoples' ratings can be used to provide recommendations.

In [83] the authors present a framework for affective classification, search and retrieval of movies where the focus is on the subjective emotional impact of movies on users. Six different emotions are determined using physiological measures [84] and assigned to movie scenes using a set of rules. Movies are also categorized according the dominant emotion (based on the amount of time an emotion was felt by a user). Each user also has access to the average emotions felt by all users for a given movie. By means of a rich visualization interface which depicts separate emotions as colors, users are able to browse movies based on their emotional similarity to one another and view emotion timelines of movies.

4.2 Demographic Recommender Systems

Demographic-based filtering extends the profile of the user to include attributes such as age, gender, ethnicity and physical location [85]. Demographics can also be learned using feature extraction combined with machine learning techniques [8]. Often it is collaborative filtering recommenders that are extended with demographic

parameters, since demographics fit well into the context of social recommendation. Recommendation to a user from a particular demographic group can then be carried out by offering them items from other users in the same demographic group. *Lifestyle Finder* is such a social demographic recommender, where in a process known as *demographic generalization*, demographic information from users are used as constraints to find one or more associated pre-computed demographic clusters [4]. The information from the matching clusters are then used as a broad demographic profile for each person, and the parameters that differentiate each cluster from the others can be used to gather additional feedback from the user to refine their profile.

4.3 Knowledge-based Recommender Systems

Knowledge-based systems, sometimes also referred to as conversational systems, are intended for seldomly-purchased items for which ratings might be scarce, such as automobiles, or simply unavailable, such as single-broadcast only television content. Instead of social ratings, knowledge-based systems harness users' domain knowledge to allow them to formulate special requirements [3]. Since a user history is not needed, they do not suffer from the same ramp-up problems experienced with other recommender approaches [86]. The knowledge-based recommendation strategy goes according to an interactive process whereby users are guided to the item of interest [3]. The user, who might be slightly undecided in the beginning, can learn their real preferences in the process as they gain more insight into the domain of interest. For example, in the case of an automobile, the user may wish to specify a requirement such as *cheaper* or *diesel engine only*. The other recommendation strategies are not suitable for addressing such requirements.

Knowledge-based systems are typically implemented as either constraint-based systems, or case-based systems. In constraint-based systems, the user specifies a set of constraints from which the system must provide a recommendation [87]. The constraints are based on the user's preferences, items' features, as well as limitations and links between the two, and items are recommended by satisfying a set of constraints for the given product features and user requirements. If the user is not satisfied with the recommendation, one or more of the constraints are altered.

In case-based systems, users use a process known as *critiquing*, where they browse an often multi-dimensional navigation item space to find suitable items [86, 88, 89]. First, an initial item is proposed to the user using some form of similarity metric. If a user is not satisfied with the initial recommendation, they can use the browsing interface to select one or more attributes that need changing, and steer the navigation in a given direction. For example, the customer might select the attribute *price*, and move in the direction *cheaper*. Sometimes, the user will have to compromise on some features in order to select other features. Hence, unlike constraint-based systems, the manner in which preferences are articulated is

more comparative. Each such browsing operation allows the system to provide a new recommendation. This process of browsing is repeated until a suitable item is found.

The back-end of case-based systems can be seen as a query interface, where requirements are encoded as queries and are updated in real time. In fact, relevance feedback using Rocchio's method can be seen as a very early form of a case-based system, where requirements are refined through iterative queries [90]. Each browsing update leads to a reduction in the item space, since now any items that no longer fulfill the new requirements are no longer taken into account. To avoid displaying critique options that lead to no possible items, critiques can also be generated dynamically by generating a set of rules that match the remaining items [91]. The rate at which items are reduced depends on whether unit or compound critiques are selected (compound critiques reduce the space more rapidly) as well as the strength of dynamic critiques (directly related to the metric *support*) to reduce the list of candidate items.

Although originally intended for once-off items such as automobiles and restaurants, case-based recommenders have shown promise for low-involvement product domains [89]. In *PickAFlick*, one of the early *FindMe* systems, domain knowledge is used to recommend candidate movies that are similar to an initially selected movie on the basis of genre, cast or director [88]. By viewing the example candidates, the user can select select additional recommendations.

In an application called *Movie Tuner* used to extend MovieLens, users critique movies with respect to a selection of tags, to give operations such as *more action* and *less violence* [92]. The movie tags are obtained from user-contributed content and the relevance of each tag to each movie is computed upfront using a hierarchical regression algorithm, allowing the system to compute similarities between items.

5 A Framework for Audio-based Recommendation of TV Content

In this section we present a proposed framework for audio-based recommendation of TV content. The full outline is shown in Figure 1. We discuss the various parts in more detail below, and also give motivations for some of the decisions taken in our approach. When referring to a block from the figure in the text, this is highlighted in **bold** font.

5.1 User Interaction

As shown in the figure, a user can interact with the system in four different ways:

1. **Speech**: The user can generate speech, from which their age and gender class, as well as their current emotion, can automatically be extracted. Once

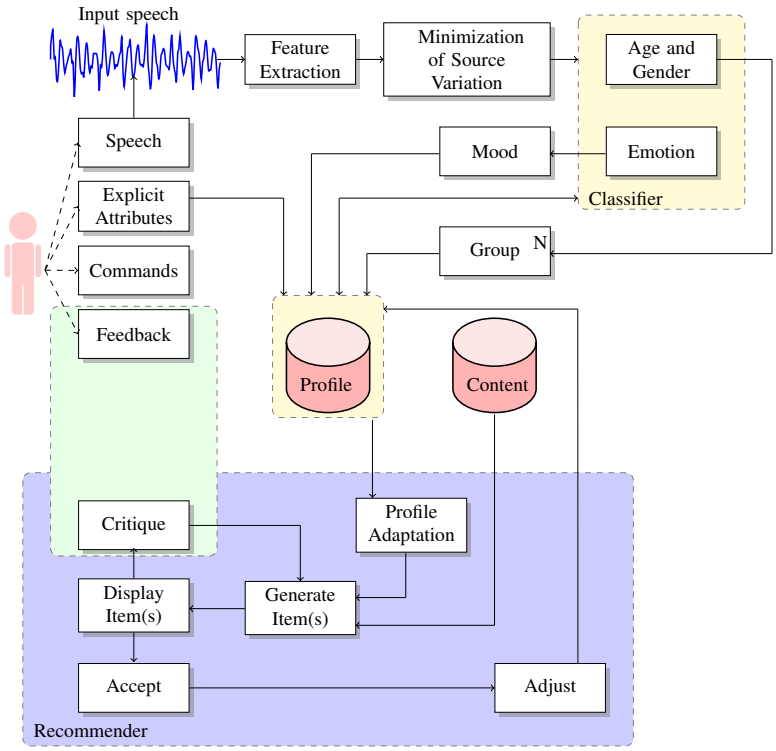


Fig. 1: A framework for recommending TV content using audio-based parameters

the speech attributes have been extracted, they can be added to the user profile.

2. **Explicit Attributes**: Before requesting a recommendation, the user can provide specific attributes, that are not amenable to being detected through speech. These parameters help to individualize the user, for example, the user's preferred genre or whether they are a high or low granularity (from an emotion perspective) person.

3. **Commands**: The user can issue a command, such as to request a recommendation for an item.

4. **Feedback**: The user can provide feedback during a recommendation session, for example when shown an item, the user can accept the item, or steer the browsing interface in the direction of their preference. Unlike specifying attributes directly, this module is tied into the recommendation session.

5.2 Feature Extraction

After carrying out preprocessing including voice activity detection on the raw audio samples [93], **feature extraction** is carried out. The type of features extracted depends very much on the type of attributes required. In the case of features for age and gender, we include acoustic as well as prosodic features, where they can complement one another. The setup is similar for emotion detection, where we combine short-term and long-term features. However, even though linguistic features have proven to be effective for emotion detection, we do not include them. The acted nature of the emotions from the chosen corpus, as will be pointed out in Section 5.7, results in identical linguistic content, and this reduces the effectiveness of such features.

5.3 Channel and Source Compensation

Since the training and test data can come from different sources, this will give rise to source variability (in addition to channel variability). **Minimization of source variability** is therefore an important preprocessing stage to ensure a more robust classification stage.

5.4 Speaker Classification Module

The **Classifier** determines to which class the incoming stream of features belongs. The number of classes depends on the application at hand (**Emotion** or **Age and Gender**), and we have selected that number to match the chosen databases.

We considered both performance and overall design when selecting a classification paradigm. For example, in one of the proposals where we detect emotions for affective offset, we selected to use just the i-vector system, although fusing the results from an SVM system could have led to better results. This was due to the requirement for fast relabeling of emotions for speech utterances, without requiring retraining of the classifier. For an SVM classifier, if we modify one single parameter of the system, the entire model needs to be retrained.

5.5 User Profile

Before recommending content, a **Profile** is created using the information extracted from the **Classifier**. Each class is assigned a probability, with the most probable class being assigned the highest probability. Given a set of user utterances and C classes, a probabilistic profile showing the membership of each class for each speaker can simply be stated as:

$$x_m = \begin{bmatrix} p_{m,1} \\ p_{m,2} \\ \vdots \\ p_{m,C} \end{bmatrix} \tag{7}$$

where $p_{m,j}$ represents the actual predicted probability for class j, $1 \leq j \leq C$. Note that $\sum_{j=1}^{C} p_{m,j} = 1$. After speaker classification, if age and gender are being detected, the **Group** sub-module might combine the demographic data from N TV users. If it is emotions being detected, the **Mood** sub-module aggregates emotion parameters over a longer time period in order to more stably determine the user's mood. In separate works, we show using a probabilistic profile how such a group profile [94, 95], or mood profile [96] can be created.

5.6 Recommendation of Items

The task of the **Recommender** is to take the user's **Profile**, as well as the available **Content**, and use that to provide personalized recommendation. **Profile Adaptation** might be required to align to the profile with the number of items to be presented. Once the profile is in the right form, the recommender system will **Generate item(s)** and then **Display item(s)** to the user. The user might be able to **Critique** the proposed item(s) if it is not the desired item, in which case new items are generated and displayed to the user. When the user is satisfied, they can then **Accept** the item. At that point, the system will use the new feedback to **Adjust** the user profile for future recommendation rounds. In certain cases, the new **Profile** can be used to update the **Classifier**.

The involvement of the user varies depending on the task at hand. For example, in a group setting, where advertisements are recommended based on the age and gender composition of a group of users, no involvement apart from speaking is required. The profile from the user (or group of users) can then be directly matched to content using a content-based approach. When browsing a user interface to find a more satisfactory item, more user involvement might be required. Also, in this type of recommendation scheme, there are multiple recommendation strategies at play: A content-based approach is used to find the initial item. After that it is left up to the user to use their knowledge to find a more appropriate item.

5.7 Chosen Databases

For age and gender, the chosen corpus was the aGender corpus [65] made available at the 2010 Interspeech Paralinguistics Challenge [72], and for emotion we chose the GEMEP corpus [52] made available at the 2013 Interspeech Paralinguistics Challenge [60]. The motivation for using the aGender corpus was the fact that it is a recent database, the availability of performance results (for comparison) and

the availability of seven age and gender classes, which provided a good match with our experiments. These seven classes are *child*, *young male*, *young female*, *adult male*, *adult female*, *senior male* and *senior female*.

The motivation for using an acted database for emotion was due to the large number of emotions available, and once again the availability of benchmark results for comparison [60, 97]. Here there are twelve emotion classes, which are *amusement*, *joy*, *interest*, *cold anger*, *hot anger*, *fear*, *anxiety*, *despair*, *sadness*, *relief*, *pride* and *pleasure*.

To aid in reproducibility of our results, we have adopted the same training, development and test splits in our work as outlined for the respective databases.

5.8 Privacy and Ethical Concerns

From an ethical standpoint, personalizing recommendations based on peoples' speech could be a difficult problem to solve. Privacy concerns are also relevant for the systems that we propose in this work. For example, demographic recommender systems are a concern with consumers since people are gradually becoming more reluctant to disclose personal information [46]. Some people feel that monitoring their viewing habits is an invasion of their privacy. The media terms commonly used to portray such products such as "George Orwell's 1984", "Big Brother", "spy TV" and "eavesdropping" are all testament to this. People are also rightfully concerned about being identified, having their personal information divulged, and being exposed to embarrassing situations, for example by displaying recommendations based on a person's viewing habits to family or friends. In a sense, these ethical concerns were a motivation for us to focus on recommendation based on the non-linguistic speech phenomena.

In 2013, an IT consultant Jason Huntley was able to determine that his LG smart TV was sending information about his viewing habits [98]. In early 2015, Samsung sparked massive media controversy when they warned consumers to refrain from private and sensitive topics in front of their latest smart TVs, as speech data could be sent to a third party (Nuance) [99]. Although the speech would only be sent at the touch of a button, it was perceived by some that conversations were being passively recorded in the background, and then re-transmitted [99].

Some law firms and consumer watchdogs agree that it might not be legal to carry out such monitoring, even if the customer is notified [100], or the customer agrees (possibly without their knowledge) to the terms and conditions. Currently, many brands of television are circumventing the ethical dilemma of collecting consumer data by simply having consumers agree to a set of terms and conditions. However, not agreeing to these terms could lead to unavailability of services. These facts have left consumers somewhere in the middle. Some consumers might be more concerned about family and friends finding out their viewing habits than some third-party, that, in a way, is unknown to them. It is clear that more cultural understanding and acceptance may be needed from both sides for this technology

to become a reality. It is encouraging however, to see such an increase in interest from the side of industry in recent years.

6 Topics of the Thesis

6.1 Linkage of papers

The main part of this thesis is a collection of six papers. The first four papers contribute to applying known speaker recognition techniques within a recommender context. The final two papers focus on the core technology of source suppression for speaker identification using i-vectors, which itself can be used to enhance speaker classification. The papers and their linkage are shown in Figure 2 below.

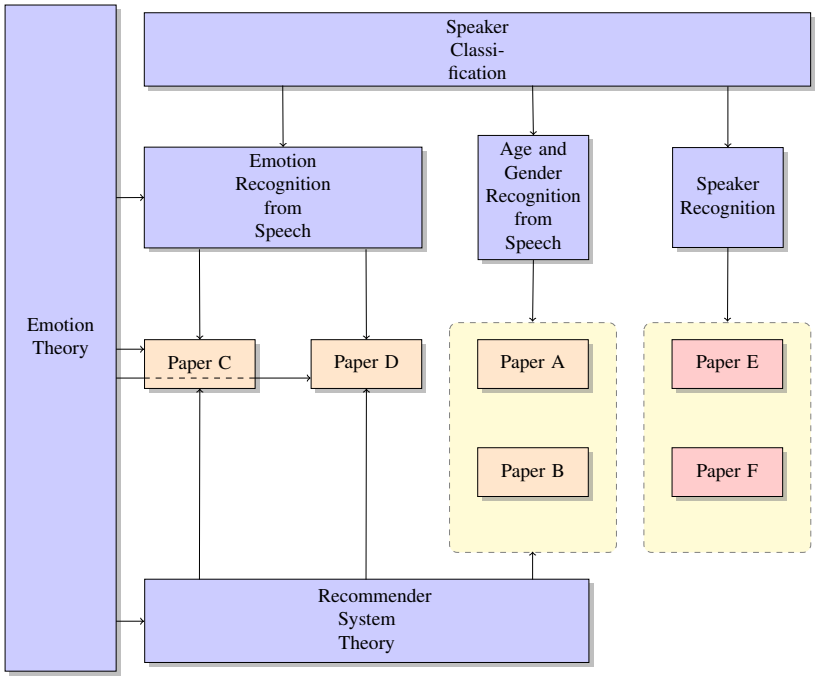


Fig. 2: Overview of contributions and their linkage within the proposed framework.

6.2 Summary of Contributions and Sub-hypotheses

1. **Paper A - Demographic Recommendation by means of Group Profile Elicitation Using Speaker Age and Gender Recognition**

   In this work, we extend demographic recommendation to a group setting, where the age and gender of a group of users is used to recommend a se-

quence of advertisement clips to that group. Assuming a set of audio utterances are available for each user, one of seven different age and gender classes is determined for the user. Combining the age and gender profile setting for all users in the group to a single profile allows a system to recommend more suitable items to the group. The sub-hypothesis is that given a particular configuration for the age and gender representation of a group of viewers, and a matching audio-based group profile for the group, that a sequence of recommended content items will receive higher ratings than if the sequence were randomly generated. This work was published in [94].

2. **Paper B - Audio-based Age and Gender Identification to Enhance the Recommendation of TV Content**

   In the aforementioned paper, the assumption was that the number of users in the group will always be the same as the number of items to be recommended, resulting in a simplified recommendation algorithm. In Paper B, we extend the work of Paper A to consider the case where the number of users is different to the number of items to be presented. An adaptation algorithm is proposed to convert the user profile to a so-called content profile, where the proportional demographic membership is retained. With this in mind, a sub-hypothesis that we investigate is that for a group of viewers in front of the TV, the demographic composition of the sequence of content items that is proposed (including the case where the number of content items does not equal the number of users) will be significantly proportional to the true demographic composition of the group of users. Another general sub-hypothesis that is investigated is that the age and gender category of each item that is proposed is significantly correlated to the user's true demographic category. The proposed system is compared to an ideal system, where group demographics are explicitly provided. This work was published in [95].

3. **Paper C - Using Audio-derived Affective Offset to Enhance TV Recommendation**

   This work is motivated by the fact that two users in the same mood will not necessarily agree on what content is applicable for a given mood setting. Given a set of utterances over a short time period, one of twelve emotion classes is determined for each time interval, and the emotion classes are condensed to form a mood profile. The mood profile is then subsequently used to propose an initial content item to the user. If the user does not find the item appealing, they are given the opportunity to continue browsing for a more suitable item. The browsing is carried out in the valence-arousal space, where the relevant user dimensions are pleasantness and intensity, respectively. Once the final item has been chosen, the affective offset between the initial and final item is recorded, stored, and used in future rounds to propose more suitable items. The sub-hypothesis is that when affective offset

is applied, the initially presented items will receive higher ratings, and for the cases where a more appealing item is desired, that a lower number of iterations, and consequently less browsing, will be needed. This work was published in [96].

4. **Paper D - Audio-based Granularity-adapted Emotion Classification**

   Evidence from the psychology domain shows that people do not possess the same discriminating abilities when it comes to detecting emotions, and therefore vary in their emotional granularity. Inspired by this fact, along with the fact that automatic detection of emotions from speech can result in certain emotions being incorrectly predicted, we propose a framework that combines machine-based and human-based emotion recognition within a recommendation context. By classifying a lower number of adapted emotion classes for high-granularity people, in principle, a larger number of content items can be presented to them, allowing them to use their strengths to make a more informed selection. The sub-hypothesis is that the adapted classes for high-granularity people will include more similar emotions, when compared to a test utterance, than single emotion classes. This paper was submitted for publication in [101].

5. **Paper E - Source-specific Informative Prior for i-Vector Extraction**

   In Papers C and D we used an i-vector system to carry out the twelve-class emotion classification - either as a standalone system or in conjunction with another system. When relying on an audio-based speech interface, there is likely to be a mismatch between the test utterance, and the target utterance of the user or class. In this work, we turn our attention to enhancing speaker recognition, in the speaker verification context, where we attempt to reduce this mismatch, known as source variation. By encoding the characteristics of each source into an informative prior, this informative prior can be used in posterior computations, on which i-vector modeling is based. In this way, source variation can be carried out directly at the i-vector extraction stage. The sub-hypothesis is that using informative priors where heterogeneous data is concerned (i.e. for multiple sources), will lead to source suppression and performance improvements. This paper was accepted for publication in [102].

6. **Paper F - Total Variability Modeling Using Source-specific Priors**

   In Paper F, we extend the work of Paper E to using informative priors not only in the i-vector extraction stage, but also in the modeling of the total variability matrix. The sub-hypothesis is that informative priors incorporated into the E-step of the EM training algorithm can lead to a better initial alignment of the total variability matrix, and can further suppress source variation. Furthermore we investigate using factor analysis to model priors

where the data might be sparse. This paper was submitted for publication in [103].

For papers E and F, the approach to using informative priors operates directly in the supervector space. To further motivate the benefit of this approach, in appendix A, we add an additional baseline that employs source-normalized and weighted LDA directly in the i-vector space, and compare the major results to this baseline[1].

7 Conclusion and Future Direction

In this thesis, we used age, gender and emotion paralinguistics from speech to enhance the way content items are recommended to TV viewers. The thesis also dealt with source variability in the context of total variability modeling, which is a core speaker recognition technology.

The first contribution (Papers A and B) described a system that utilized the age and gender of a group of viewers to present a sequence of suitable advertisements to them, and that took into account the proportional demographics of the group. The effectiveness of the system was evaluated by comparing the ratings given for a sequence of recommended adverts using age and gender, as opposed to a sequence of random adverts. It was found that the sequence of recommended items received higher ratings.

The second contribution (Paper C) was the design of a system that determined a user's mood from past emotions, and mapped this mood to a suitable content item. If the user was not happy with the item that was presented, they were given the chance to critique the item and propose an alternative. The affective offset was determined and recorded, and used for future recommendation rounds. The effectiveness of the system was determined in two ways. Firstly, ratings were compared for the items presented to users before and after applying affective offset, where it was found that after applying affective offset, that the initially recommended items received higher ratings. Secondly, the number of iterations (critiquing cycles) needed to find a liked item was compared before and after applying affective offset. Here it was found that a lower number of cycles was needed to find a desired item once affective offset had been applied.

The third contribution (Paper D) applied the emotional granularity of individuals to enhance classification of emotions, with the ultimate goal of increasing the potential similarity between the emotion of the user's speech and a content item. This was done by mapping each person's valence focus and arousal focus to a set of adapted classes, and using these instead for emotion classification. The system was evaluated by comparing each subject's proposed granularity-adapted

[1] As Paper F is still undergoing peer review, the additional baseline will be incorporated into Paper F in the next revision.

emotion classifier to a standard baseline classifier. Here it was found that for high-granularity subjects, that there was potential for including emotions (and hence mapped items) that were closer similarity-wise to the ground-truth of the spoken emotion.

The final contribution (Papers E and F) introduced the notion of using informative priors to reduce source variability in a speaker verification scenario. An informative prior, one for each source, was estimated and used instead of the usual informative prior when computing the posterior distribution. This posterior distribution is central to the computation of the total variability matrix as well as determining i-vectors for speech utterances. The system was evaluated using the NIST 2008 and NIST 2010 speaker recognition evaluation (SRE) dataset. Here it was found that informative priors led to performance improvements in the majority of cases for both evaluations.

There are many avenues for future work, and the proposals we state here are just the tip of the iceberg. Future work might investigate the use of keyword spotting techniques, where the actual content of speech can be used to provide content or knowledge-based recommendation. Tags obtained from keyword spotting could also be used to extend collaborative filtering techniques. A possible use-case might be to determine which tags of a given predefined set are detected, and then to use these tags (or the frequency of usage of each tag) to provide implicit ratings for content. Future work might even look at additional paralinguistic phenomena, and how these could be used to provide recommendation. For example, the automatic detection of a person's intoxication level, which has been a hot research topic in recent years, but mostly of interest to the police force, might be applied in the entertainment domain to suggest more appropriate content (perhaps calmer content) to such viewers. Other interesting phenomena include recommending calm programs to stressed viewers and suggesting more regional content to hotel room guests by detecting their dialect. From a performance perspective, it would be interesting to see to what extent the accuracy of a speaker or speech recognition algorithm could affect the accuracy of recommendations. In more concrete terms, one might envision a system where ratings are communicated through a speech interface, and where the identity of a given rating is provided by a speaker recognition algorithm. It would be interesting here to investigate what the effects are of mistaking a giver user for another user, how this would affect the mean absolute error (MAE) of a collaborative filtering algorithm, and how one might address this. From the perspective of user acceptance, additional work might look at how to gain user trust within the audio-based recommendation framework. From the perspective of ethics and privacy, one might look at how recommendation based on spoken content can be achieved in an affective way, while maintaining the anonymity of consumers.

References

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 6, pp. 734–749, 2005.

[2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[3] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: An introduction*. Cambridge University Press, 2010.

[4] B. Krulwich, "Lifestyle finder: Intelligent user profiling using large-scale demographic data," *AI magazine*, vol. 18, no. 2, p. 37, 1997.

[5] S. H. Hsu, M.-H. Wen, H.-C. Lin, C.-C. Lee, and C.-H. Lee, "AIMED - a personalized TV recommendation system," *Lecture Notes in Computer Science*, vol. 4471, pp. 166–174, 2007.

[6] A. B. Barragáns-Martíneza, E. Costa-Montenegroa, J. C. Burguilloa, M. Rey-Lópezb, F. A. Mikic-Fontea, and A. Peleteiroa, "A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition," *Information Sciences*, vol. 180, pp. 4290–4311, 2010.

[7] M. Z. Bjelica, "Unobtrusive relevance feedback for personalized TV program guides," *IEEE Transactions on Consumer Electronics*, vol. 57, pp. 658–663, 2011.

[8] M. J. Pazzani, "A framework for collaborative, content-based and demographic filtering," *Artificial Intelligence Review*, vol. 13, no. 5-6, pp. 393–408, 1999.

[9] D. Bonnefoy, M. Bouzid, N. Lhuillier, and K. Mercer, ""More Like This" or "Not for Me": Delivering personalised recommendations in multi-user environments," *Lecture Notes in Computer Science*, vol. 4511, pp. 87–96, 2007.

[10] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

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

[12] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[13] A. Dempster, N. Laird, and D. B. Rubin, "Maximum likelihood estimation via the EM algorithm," *J. of the Stat. Roy. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[14] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[15] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.

[16] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[17] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[18] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[20] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *Proceedings of Interspeech, Firenze, Italy*, pp. 861–864, 2011.

[21] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech." in *Odyssey*, 2010, p. 6.

[22] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition." in *Interspeech*, 2006, pp. 1471–1474.

[23] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[24] S. J. Prince, *Computer vision: models, learning, and inference*. Cambridge University Press, 2012.

[25] Y. Jiang, K. A. Lee, and L. Wang, "PLDA in the i-supervector space for text-independent speaker verification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–13, 2014.

[26] J. A. Villalba and E. Lleida, "Handling i-vectors from different recording conditions using multi-channel simplified plda in speaker recognition." in *ICASSP*, 2013, pp. 6763–6767.

[27] National Institute of Standards and Technology, "The NIST 2008 SRE Evaluation Plan," 2008. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2008/

[28] ——, "The NIST 2010 SRE Evaluation Plan," 2010. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/

[29] M. McLaren and D. Van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5456–5459.

[30] ——, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 755–766, 2012.

[31] N. Dehak, Z. N. Karam, D. A. Reynolds, R. Dehak, W. M. Campbell, and J. R. Glass, "A channel-blind system for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4536–4539.

[32] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[33] Y. Lei, L. Ferrer, M. McLaren, and N. Scheffer, "A deep neural network speaker verification system targeting microphone speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[34] C. Müller, *Speaker Classification I*. Springer, 2007.

[35] R. Cowie, N. Sussman, and A. Ben-Ze'ev, "Emotion: Concepts and definitions," in *Emotion-Oriented Systems*. Springer, 2011, pp. 9–30.

[36] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[37] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[38] J. A. Russel, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1170, 1980.

[39] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[40] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[41] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143–154, 2005.

[42] L. A. Feldman, "Valence focus and arousal focus: Individual differences in the structure of affective experience." *Journal of personality and social psychology*, vol. 69, no. 1, p. 153, 1995.

[43] L. F. Barrett, "Feelings or words? understanding the content in self-report ratings of experienced emotion." *Journal of personality and social psychology*, vol. 87, no. 2, p. 266, 2004.

[44] P. Winoto and T. Y. Tang, "The role of user mood in movie recommendations," *Expert Systems with Applications*, vol. 37, no. 8, pp. 6086–6092, 2010.

[45] M. Tkalcic, A. Kosir, and J. Tasic, "Affective recommender systems: the role of emotions in recommender systems," in *Proceedings of the 5th ACM conference on recommender systems*, 2011, pp. 9–13.

[46] A. T. Ho, I. L. Menezes, and Y. Tagmouti, "E-MRS: Emotion-based movie recommender system," in *Proceedings of IADIS e-Commerce Conference. USA: University of Washington Both-ell*, 2006, pp. 1–8.

[47] C. Peter and A. Herbon, "Emotion representation and physiology assignments in digital systems," *Interacting with Computers*, vol. 18, no. 2, pp. 139–170, 2006.

[48] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 10, pp. 1175–1191, 2001.

[49] Z. Zeng, Y. Hu, G. I. Roisman, Z. Wen, Y. Fu, and T. S. Huang, "Audio-visual spontaneous emotion recognition," in *Artifical Intelligence for Human Computing*. Springer, 2007, pp. 72–90.

[50] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A dimensional approach to emotion recognition of speech from movies," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 65–68.

[51] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech." in *Interspeech*, 2005, pp. 1517–1520.

[52] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception." *Emotion*, vol. 12, no. 5, p. 1161, 2012.

[53] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a danish emotional speech database." in *Eurospeech*, 1997.

[54] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. J. Russell, and M. Wong, ""You Stupid Tin Box"-Children interacting with the AIBO robot: A cross-linguistic emotional speech corpus." in *LREC*, 2004, pp. 171–174.

[55] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 865–868.

[56] R. E. Madsen, J. Larsen, and L. K. Hansen, "Part-of-speech enhanced context recognition," in *Machine Learning for Signal Processing, 2004. Proceedings of the 2004 14th IEEE Signal Processing Society Workshop*. IEEE, 2004, pp. 635–643.

[57] M. Kockmann, L. Burget *et al.*, "Application of speaker-and language identification state-of-the-art techniques for emotion recognition," *Speech Communication*, vol. 53, no. 9, pp. 1172–1185, 2011.

[58] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5688–5691.

[59] R. Xia and Y. Liu, "Using I-vector space model for emotion recognition." in *INTERSPEECH*, 2012, pp. 2230–2233.

[60] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 148–152.

[61] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *INTERSPEECH*, vol. 2009, 2009, pp. 312–315.

[62] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals." in *INTERSPEECH*. Citeseer, 2007, pp. 2253–2256.

[63] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search." in *Interspeech*. Citeseer, 2009, pp. 1151–1154.

[64] H. Meinedo, "Audio pre-processing and speech recognition for broadcast news," *Universidade Técnica de Lisboa, Diss*, 2008.

[65] F. Burkhardt, M. Ekert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," *Proc. 7th International Conference on Language Resources and Evaluation (LREC)*, pp. 1562–1565, 2010.

[66] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. J. Russell, S. Steidl, and M. Wong, "The PF-STAR children's speech corpus." in *INTERSPEECH*, 2005, pp. 2761–2764.

[67] T. I. Meinedo H., "Age and gender detection in the I-DASH project," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, 2011.

[68] H. Meinedo and I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features." in *INTERSPEECH*, 2010, pp. 2818–2821.

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

[70] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. G. Bauer *et al.*, "Comparison of four approaches to age and gender recognition for telephone applications," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1089.

[71] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoutic and prosodic level information fusion," *Computer Speech and Language*, vol. 27, pp. 151–167, 2012.

[72] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge." in *INTERSPEECH*, 2010, pp. 2794–2797.

[73] H. Höge, C. Draxler, H. van den Heuvel, F. T. Johansen, E. Sanders, and H. S. Tropf, "Speechdat multilingual speech databases for teleservices: across the finish line." in *EUROSPEECH*, vol. 99, 1999, pp. 5–9.

[74] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[75] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech communication*, vol. 25, no. 1, pp. 117–132, 1998.

[76] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[77] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[78] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 43–52.

[79] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[80] D. Watson and L. A. Clark, "The PANAS-X: Manual for the positive and negative affect schedule-expanded form," 1999.

[81] M. Garden and G. Dudek, "Semantic feedback for hybrid recommendations in recommendz," in *e-Technology, e-Commerce and e-Service, 2005. EEE'05. Proceedings. The 2005 IEEE International Conference on*. IEEE, 2005, pp. 754–759.

[82] E. Amolochitis, I. T. Christou, and Z.-H. Tan, "Implementing a commercial-strength parallel hybrid movie recommendation engine," *Intelligent Systems, IEEE*, vol. 29, no. 2, pp. 92–96, 2014.

[83] E. Oliveira, P. Martins, and T. Chambel, "Accessing movies based on emotional impact," *Multimedia systems*, vol. 19, no. 6, pp. 559–576, 2013.

[84] E. Oliveira, M. Benovoy, N. Ribeiro, and T. Chambel, "Towards emotional interaction: using movies to automatically learn users' emotional states," in *Human-Computer Interaction–INTERACT 2011*. Springer, 2011, pp. 152–161.

[85] R. Burke, "Hybrid recommender systems: Survey and experiments," *User modeling and user-adapted interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[86] ——, "Knowledge-based recommender systems," *Encyclopedia of Library and Information Science: Volume 69-Supplement 32*, vol. 69, no. 32, pp. 180–200, 2000.

[87] A. Felfernig and R. Burke, "Constraint-based recommender systems: technologies and research issues," in *Proceedings of the 10th international conference on Electronic commerce*. ACM, 2008, p. 3.

[88] R. D. Burke, K. J. Hammond, and B. Yound, "The FindMe approach to assisted browsing," *IEEE Expert*, vol. 12, no. 4, pp. 32–40, 1997.

[89] L. Chen and P. Pu, "Critiquing-based recommenders: survey and emerging trends," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 125–150, 2012.

[90] J. J. Rocchio, "Relevance feedback in information retrieval," 1971.

[91] J. Reilly, K. McCarthy, L. McGinty, and B. Smyth, "Dynamic critiquing," in *Advances in Case-Based Reasoning*. Springer, 2004, pp. 763–777.

[92] J. Vig, S. Sen, and J. Riedl, "Navigating the tag genome," in *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, 2011, pp. 93–102.

[93] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 798–807, 2010.

[94] S. E. Shepstone, Z.-H. Tan, and S. H. Jensen, "Demographic recommendation by means of group profile elicitation using speaker age and gender detection," *Proceedings of Interspeech, Lyon, France*, pp. 2027–2031, 2013.

[95] ——, "Audio-based age and gender identification to enhance the recommendation of TV content," *Consumer Electronics, IEEE Transactions on*, vol. 59, no. 3, pp. 721–729, 2013.

[96] S. Shepstone, Z.-H. Tan, and S. H. Jensen, "Using audio-derived affective offset to enhance TV recommendation," *Multimedia, IEEE Transactions on*, vol. 16, no. 7, pp. 1999–2014, 2014.

[97] G. Gosztolya, R. Busa-Fekete, and L. Tóth, "Detecting autism, emotions and social signals using AdaBoost," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2013, pp. 1–5.

[98] S. Griffiths, "Is your smart tv spying on your family? investigation reveals how much personal data television sets know about viewers," *Daily Mail*, August 2014. [Online]. Available: www.dailymail.co.uk/sciencetech/article-2737708

[99] J. Hamill, "Samsung spy telly scandal erupts after firm admits its television will record your "personal and sensitive" conversations," *Mirror*, February 2015. [Online]. Available: http://www.mirror.co.uk/news/technology-science/technology/samsung-spy-telly-scandal-erupts-5131842

[100] A. Ionescu, "Samsung's technology contravenes danish law," *The Copenhagen Post*, February 2015. [Online]. Available: http://cphpost.dk/news/samsungs-technology-contravenes-danish-law.12597.html

[101] S. E. Shepstone, Z.-H. Tan, and S. H. Jensen, "Audio-based granularity-adapted emotion classification," *Affective Computing, IEEE Transactions on*, Submitted Paper.

[102] S. E. Shepstone, K. A. Lee, H. Li, Z.-H. Tan, and S. H. Jensen, "Source specific informative prior for i-vector extraction," in *Acoustics Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, Accepted Paper.

[103] ——, "Total variability modeling using source-specific priors," *Audio, Speech, and Language Processing, IEEE Transactions on*, Submitted Paper.

Part II

Papers

Paper A

Demographic Recommendation by means of Group Profile Elicitation Using Speaker Age and Gender Recognition

Sven Ewan Shepstone, Zheng-Hua Tan and Søren Holdt Jensen

The paper was published in the
Proceedings of Interspeech 2013 pp. 2027–2031, 2013.

The layout has been revised.

Abstract

In this paper we show a new method of using automatic age and gender recognition to recommend a sequence of multimedia items to a home TV audience comprising multiple viewers. Instead of relying on explicitly provided demographic data for each user, we define an audio-based demographic group profile that captures the age and gender for all members of the audience. A 7-class age and gender classifier employing a fusion of acoustic and prosodic features determines the probability of each speaker belonging to each class. The information for all speakers is then combined to form the group profile, which itself is the input to a recommender system. The recommender system finds the content items whose demographics best match the group profile. We tested the effectiveness of the system for several typical home audience configurations. In a survey, users were given a configuration and asked to rate a set of advertisements on how well each advertisement matched the configuration. Unbeknown to the subjects, half of the adverts were recommended using the derived audio demographics and the other half were randomly chosen. The recommended adverts received a significantly higher median rating of 7.75, as opposed to 4.25 for the randomly selected adverts.

1 Introduction

This paper shows how a state-of-the-art age and gender classifier can be leveraged to power a recommender system for selecting TV content. Instead of basing the age and gender profile needed for recommendation on manually provided data or usage patterns, we propose using audio analysis methods instead.

The detection of age and gender is a complicated task and has received a lot of research interest recently. Typically, the age and gender of speakers are identified by means of Gaussian mixture models, multilayer perceptrons, hidden Markov models and/or support vector machines [1], [2]. In particular, modern age and gender classification results are making it more and more feasible to use on-the-fly demographic classification for recommendation purposes. The state-of-the-art accuracy of gender-only classifiers is roughly 30 % higher than that of age detection [3]. The same work shows that a system using automatic speaker recognition using a fusion of acoustic and prosodic features was able to achieve an accuracy of 85.0 % for the gender classification task, 52.0 % for the age classification task and an accuracy of 50.3 % for the combined age and gender classification task [3].

What is interesting to note is that the largest confusion occurs between speakers of the same sex (e.g. young males, adult males and senior males) and between children and young females. While there is still room for future improvement, we believe that there is a strong basis for recommendation, since the effect of overlapping confusion classes could well be ameliorated by soft preference and

market boundaries. For example, with respect to short advertisement clips, there are many products that would appeal to both young males and adult males, or to both children and young females, thus canceling out some of the effects of the confusion overlap between these classes.

Collaborative recommender systems are the most widespread recommender systems in use today and rely on a large user base of ratings to make recommendations. Essentially, these systems work by correlating the feedback rating of a user for a specific item with that of other users for the same item, to make recommendations for a new item that is unknown to the user (but that the others rated) [4]. However, with home set-top boxes there is no easy way to exchange the user ratings, with the result that for these types of systems, a content-based approach is more applicable [5].

Content-based recommenders can determine similarities directly between content items and a given user profile, provided the user profile can be extracted, and there exists suitable meta-data for content items[1]. However, the need for a user profile implies that the profile must either be explicitly provided, for example by means of a questionnaire when registering a set top box [6], or implicitly, by building the profile by monitoring usage patterns [5]. A bigger problem, however, is when multiple consumers share a single device, such as a home television, but each has their own user profile and tastes [7]. This occurs often with home game playing and movie watching, where typically only one username or profile is utilized.

Our contribution in this paper is a novel method of using audio analysis techniques to extract the parameters needed for constructing a group profile for recommendation. This is in contrast to traditional methods of using user questionnaires, usage data or ratings to collect the viewers' data. We focus primarily on age and gender in this study, and utilize an age and gender classifier to provide a group demographic profile for communal TV viewing. We test the hypothesis that given a particular home viewer configuration, and given a group profile derived using an audio analysis of each member of the configuration (audience), that the recommended items (advertisements) will receive higher ratings from users, than if the content items were randomly selected, thus indicating a closer match to the viewer configuration[2].

The remainder of the paper is as follows: Section 2 introduces the notion of a demographics-based audio group profile. We then discuss adapting the group profile to make it usable for recommendation. Section 4 presents the home viewer configuration used in this study, and the audio classifier that transforms a viewer configuration to a group profile. We then discuss experimental work and the surveys that were conducted. Finally we present our results and draw conclusions.

[1] Collaborative systems and content systems are often deployed in a hybrid configuration to take advantage of their strengths.

[2] We do not evaluate the system using prediction error, since there is no ground truth (all ads rated for every group viewer configuration).

2 Extracting the Audio Group Profile

Solving the "Who is sitting in front of the TV?" problem is challenging and has yet to be researched fully. A typical system could be realized as follows: The audio from several microphone pickups in a room could be applied to an independent component analysis algorithm that separates the background TV audio (if any) from the users' speech [8]. Speaker diarization is used on the speech part to separate speaker utterances of different people from one another, and to determine the number of speakers present [9], [10]. The speaker utterances from each speaker can then be classified according to age and gender, which in turn can be used to construct a group profile. Due to the limited accuracy of current state-of-the-art age and gender systems, it is important to note that each speaker, regardless of their age and gender class, will to some extent be a member of all defined age and gender classes. In this study, motivated by the corpus that was used for training our classifier [11] and by recent works [3], [1], we base our study on seven such classes.

The user profile for each speaker m, generated over a set of utterances for that speaker, can be modeled by:

$$x_m = \begin{bmatrix} p_{m,1} \\ p_{m,2} \\ \vdots \\ p_{m,C} \end{bmatrix} \tag{A.1}$$

where $p_{m,j}$ simply represents the actual predicted probability for class j, $1 \leq j \leq C$. The more utterances that can be collected, the better the classification accuracy.

For a set of M users, we then define a group profile as:

$$X_G = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,C} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \cdots & p_{M,C} \end{pmatrix} \tag{A.2}$$

3 Matching and Recommendation

The matching problem can be stated as optimizing the match between the group profile X_G, obtained by classifying a set of utterances for each speaker, and the sequence of content items (ads) that the viewers will see. When the number of users M is equal to the number of items N we allocate *one* item per viewer, thus allowing each viewer to see a content item of their liking.

When $M \neq N$ (N might be fixed, due to e.g. the length of an ad break) there is no longer a 1-1 mapping between users and items. In this case we perform what we

refer to as group profile adaptation. This entails converting the group profile X_G, which represents M users, to a new profile Y_G, which represents N pseudo-users, and where N is now equal to the number of items to present. This means that for each class in the original group profile, we determine the proportional membership of each user to that class. For example, assume a 2-user group profile that must be extended to 3-pseudo users. For the first class (Child), we find that the first user has a 80 % membership of the class (leaving only 20 % to all other classes), while the second user has a 40 % membership of the class (leaving 60 % to all other classes). For 3 pseudo-users, we split the pseudo-user space up into 3 equally-sized portions. The first pseudo-user overlaps completely with the 1st user - hence it receives an 80 % membership. The 2nd pseudo-user overlaps $50 - 33.3 = 16.7$ % with the 1st user and $66.6 - 50 = 16.7$ % with the 2nd user. The membership for this pseudo-user is then proportionally calculated as $\frac{80*16.7+40*16.7}{16.7+16.7}\% = 60$ %. Finally since the third pseudo-user overlaps completely with the 2nd user, we just assign the same membership of user 2 to the third pseudo-user, i.e. 40 %. Note that when $M = N$ then $Y_G = X_G$.

Now for a given content item

$$c_n = \begin{bmatrix} p_{n,1} \\ p_{n,2} \\ \vdots \\ p_{n,C} \end{bmatrix} \tag{A.3}$$

which has a predefined age and gender profile, the strength of the match for each user-item pair is then simply computed as:

$$Match_{n,n} = Y_G(n, *) * c_n \tag{A.4}$$

To perform the actual matching we use a modified form of genetic algorithm, proposed previously for providing itinerary-based recommendations [12]. Genetic Algorithms are established computational methods that conduct their searches based on natural selection and genetics, and use the concepts of chromosomes, populations, selection, crossover and mutation [13].

Upon initialization, the algorithm selects k chromosomes, each containing N randomly-chosen ads. The strength of each chromosome (how well it matches the adapted group profile) is then computed by taking the sum of content-item matches, with each match computed as shown in Equation A.4 above. With each iteration of the algorithm, the chromosome with the poorest match to the adapted group profile is discarded, and replaced with a new genetically-spawned sequence.

For our experiments, the ad selection process was as follows: We first initialized our genetic algorithm with $k = 50$ chromosomes of 5 ads each. The ads were taken from a central pool of 200 ads and it was not possible for an ad to appear twice within a given chromosome. We then ran 500 iterations of genetic selection, and

selected the sequence with the strongest match likelihood as the sequence of ads to be recommended.

4 Age and Gender Audio Classification

4.1 Viewer Configuration Profile

To emulate a group containing several viewers of varying demographics, we define a viewer configuration profile. To select which viewer configurations to use, we turned to Statistics Denmark [14], which records comprehensive statistics on the composition of Danish households. Here we could see that 23.8 % of the population live alone, 38.7 % live with one other person, 14.3 % belong to a family of three, 14.6 % belong to a family of four and 5.5 % belong to a family of five. From these figures, we based our viewer configurations on families of two, three and four persons, where the bulk of the distribution lies.

Now just for the two-person households, children and youngsters don't feature much, and only comprise 2.8 % and 2.1 % of households, respectively. In contrast, 37.6 % of households contain adults and 57.5 % have seniors, giving configurations 1 and 2 in Table A.1 below.

Looking at children and youth from just the three- and four-person households, we note that for children, 30.1 % are part of three-person families, but that 69.4 % (more than double) are part of four-person families. For the youth category, 40.1 % of youths belong to three-person families whereas 59.9 % of youths belong to four-person families. Thus it is evident that children and youths should feature fairly strongly in our chosen configurations. From this, we construct configurations 3, 4, 5, 6, 7, 8 and 9 shown in Table A.1 below[3].

Finally we examined statistics on the number of seniors (≥ 55) with children and/or youngsters living at home. We found that there were twice as many seniors with two children living at home (15657 people) than seniors with only one child living at home (7302), giving the last two configurations.

Profile No	1	2	3	4
Profile	AM+AF	SM+SF	C+C	C+YM
Profile No	5	6	7	8
Profile	C+YF	C+AM	C+AF	C+C+AM
Profile No	9	10	11	
Profile	C+C+AF	C+SM	C+SF	

Table A.1: TV viewer configuration. C=Child, YM=Young Male, YF=Young Female, AM=Adult Male, AF=Adult Female, SM=Senior Male, SF=Senior Female

[3] In this study we focus on 2 and 3 people at a time in front of the TV.

This gives a total of 11 configurations. For each configurations that was presented (explained below), we broke it up into its constituent parts, i.e. individual speakers, and for each speaker, connected them to real speaker utterances.

4.2 Dataset

The speaker utterances used for classification were taken from the aGender corpus, which was supplied to participants in the InterSpeech 2010 Paralinguistic Challenge to enhance the development of age and gender algorithms [11]. The training part of the dataset contains 32527 utterances from 472 speakers, the development part contains 20549 utterances from 300 speakers and the testing part contains 17332 utterances. It comprises 4 age classes: children (7-14 years), young people (15-24 years), adults (25-54 years) and seniors ($\geq$ 55 years), and 3 gender classes: children[4], males and females. In more recent work, the age boundaries are slightly different, i.e. children ($\leq$ 13 years), young people (14-19 years), adults (20-54 years) and seniors ($\geq$ 55 years) [3]. We chose to use the latter age boundaries from the recent work.[5]

4.3 Speaker Classification

For each speaker from the viewer configuration profile we randomly selected a speaker with the matching class in the evaluation portion of the aGender dataset. To represent this speaker we pooled together the selected speaker's utterances to form a contiguous segment. Each speech segment was then submitted for classification, to determine its class. The speaker results were then combined to form the group profile $\bar{X_G}$ from above.

For classification we employed a hybrid system, where each age and gender class is modeled separately. Both acoustic and prosodic features are modeled, with fusion of acoustic and prosodic features occurring at the utterance level.

The GMM baseline was constructed using the well-known UBM-GMM approach [15]. After voice activity detection [16], feature extraction was performed using 13-dimensional MFCCs (including C0, 1st and 2nd derivative), to give 39 coefficients per 25 ms frame (15 ms overlap). We then trained a 512-component GMM UBM using all the training data from the aGender corpus. Following this, 7 speaker models were adapted from the UBM using the training data from each class. For the adaptation process, we used a relevance ratio of 12. The accuracy for the acoustic sub-system for all classes was 49.9 %.

To model the prosody features we used the prosody baseline referred to as System 7 in a previous work [3], and which models prosody features at the syllable

[4]Children are classed as their own gender since males are indistinguishable from females at that age.

[5]The original aGender age boundaries were chosen solely on the basis of marketing aspects, and not on any physiological aspects.

level instead of the frame level. The syllable boundaries are determined as follows: For each utterance, all frames are marked as voiced or unvoiced (unvoiced where the pitch is undefined) and all unvoiced frames are discarded. For the remaining frames, the normalized energy contour is used as a key to determining the syllable boundaries, where valleys in the contour indicate the start of a new syllable.

The prosody features modeled for each syllable are contours of pitch, energy, formants, syllable duration and spectral harmonic energy (obtained from the power spectrum at harmonics of F0). We used the Praat package [17] to extract pitch and energy features from each utterance and Matlab to compute the spectral harmonic energy. After applying time scale normalization for the interval -1 to 1, the contours were then modeled as sixth-order Legendre polynomials, meaning that instead of an entire contour, only six coefficients need to be stored [18]. We then trained 7 512-component GMM models with the prosody features, one for each class. The accuracy for the prosodic sub-system for all classes was 42.0 %.

We then combined the two acoustic and prosodic sub-systems together in a hybrid system using weighted summation-based fusion [3] of the sub-system results. We tested our hybrid classifier model on the entire development data set, where we achieved an accuracy on the combined system of 50.0 %. As a comparison, another work using seven individual sub-systems was able to attain an accuracy of 50.3 % [3]. A more detailed breakdown of the 2 classifiers is shown in Table A.2 below.

	C	YM	YF	AM	AF	SM	SF
C	**69.6**	3.4	16.2	1.7	4.4	1.3	3.5
	61.0	7.5	16.9	2.0	4.9	1.0	6.7
YM	1.6	44.8	1.3	27.4	0.3	19.7	4.9
	0.3	**49.4**	0.8	21.9	1.0	23.5	3.2
YF	18.7	2.2	49.9	1.3	21.6	0.5	5.7
	16.4	0.8	**57.1**	0.3	15.8	0.6	9.0
AM	2.3	20.7	0.3	**47.8**	1.3	25.1	2.5
	0.1	29.2	0.0	27.1	1.1	40.5	2.2
AF	10.4	3.5	21.1	1.9	**40.2**	1.0	21.9
	5.5	1.8	26.6	0.4	33.8	0.6	31.3
SM	2.5	14.5	0.2	23.6	0.5	55.9	2.8
	0.2	11.5	0.1	16.2	0.2	**69.7**	2.0
SF	10.5	4.7	11.6	2.1	24.9	4.3	41.9
	7.1	1.5	11.4	0.9	22.9	2.2	**53.9**

Table A.2: Confusion matrix for seven-class Age and Gender Classifier. Shaded entries are the results for our classifier (two sub-systems; overall accuracy 50.01). Non-shaded entries are the results of a recent work (seven sub-systems; overall accuracy 50.3). **Bold** typeface shows the better score of the two systems.

5 Experimental Work

The advertisement corpus used in this paper has 24 categories of ads and was provided to us courtesy of TV2, a Danish public-service television broadcaster. To be able to match advertisements with the group profile discussed above, we conducted a pre-survey to annotate each ad with an age and gender profile. We took a random subset of ads from each category, giving a total of 200 commercials, which we then split into four separate groups. For each group of 50 ads, three subjects were asked to rate all 50 commercials, on the basis of how well they thought each ad matched all seven age and gender classes. The scale used was the standard 1-5 Likert scale (1 not-relevant and 5 most relevant). For each ad rated by three separate people, we took the median rating for each class as the official rating for the advertisement. Table A.3 shows a sample selection of ads, with their corresponding median ratings.

Short Description of Ad	C	YM	YF	AM	AF	SM	SF
Women's sandals	1	1	5	2	5	1	4
Cleaning Agent	1	1	4	3	5	1	5
Lift Chair	1	1	1	1	1	5	5
Chewing Gum	1	5	4	4	4	4	4
Dating Site	1	1	1	5	5	2	2
Hair Product	1	3	5	1	5	1	5
Chocolate Easter Egg	5	1	1	3	3	1	2
Building Blocks	5	1	1	2	5	2	3

Table A.3: Selected ads with accompanying ratings.

To test the effectiveness of using the acquired audio group profile in our recommender, we conducted another survey where subjects were shown a set of home viewer configurations, and for each configuration, asked to rate a set of 10 advertisements. A different set of advertisements was used for each round. For each set shown, five of the ads were obtained by using the genetic algorithm approach and the other five were randomly selected (without replacement) from an initial pool of 200. The set was then shuffled before being presented for recommendation. Subjects were not told that five of the ads for the given slot had been randomly selected, thus giving them no way of knowing which of the ads had been recommended. They were then asked to rate each ad on a scale of 1-10 (1 completely irrelevant and 10 most relevant), on the basis of the ad appealing to *any* of the members of the home viewer configuration. For example, if the subject thought the ad appealed highly to children, and the *Child* category was part of the configuration, then the ad would receive a higher rating. A 10-point scale was used to ensure that subjects took a non-neutral stance when rating.

We used 12 subjects for our evaluation. Since it was not possible time-wise for

each subject to rate all 11 proposed viewer configurations, we split the configurations into 3 groups. The first four subjects were therefore asked to evaluate the first four group viewer configurations, the second four subjects were asked to rate advertisements for the second four configurations, and the last four subjects were asked to rate advertisements for the last three configurations.

6 Results

We now look at the results that were obtained. Table A.4 shows two median ratings for each user of the survey. The first rating was taken as the median of all ratings performed for the user on the randomly selected ads, whereas the second rating was taken as the median of all ratings for the recommended ads.

Test Subject	Random	Recommended
1	7	9.5
2	2.5	7
3	10	8
4	7	9
5	4.5	10
6	3	5
7	4	5
8	4.5	7.5
9	4	8
10	4	10
11	2	7
12	8	7

Table A.4: Average ratings for the 12 users, taken for the random case and recommended case. Average for each user taken using the median.

From the averages in Table A.4 we see that the recommended ads obtained consistently higher ratings than the random ads. Only 2 of the users (users 3 and 12) returned an average rating for the random ads that was higher than the recommended ads.

To test the statistical significance of the recommended ads receiving higher ratings, we let x represent all samples corresponding to the median ratings of all users for the random ads and y represent samples corresponding to the median ratings of all users for recommended ads, and test the null hypotheses that $y - x$ comes from a distribution of zero median. Treating the rating scales as ordinal, we use the 2-sided Wilcoxen Signed Rank test to test for significance. We find with a z-score $z = -2.628$ and $p < .01$ that there is a significant increase in the median rating for each test group, thus disproving the null hypothesis. Indeed, from the table above, the user ratings for the recommended group have a median

of 7.75, which was significantly higher than the ratings for the random group, with a median of 4.25. We also compute the effect size using Pearson's correlation coefficient $r = \frac{Z}{\sqrt{N}}$, where Z is the z-score from above and $N = 24$ is the number of observations, and find it to be $r = -0.535$. Since the absolute value is above Cohen's benchmark of 0.5, we can conclude that using the age-and-gender analysis approach has a large effect on the user ratings.

7 Conclusion

This paper showed how an age and gender classifier using mixed acoustic and prosodic features can be used to elicit a demographic group profile from a given audience, and how this can be used to provide recommendations. The classifier we built delivered comparable results to the state-of-the-art and showed that there is a basis for recommendation, even with large gaps of confusion between classes. We showed that ratings for adverts recommended using the age and gender data were significantly higher than ratings for randomly selected adverts.

Acknowledgment

The author wishes to thank Bang and Olufsen A/S for sponsoring this research, TV2 Denmark for providing the video TV commercials and Felix Burkhardt for supplying the aGender corpus.

References

[1] T. I. Meinedo H., "Age and gender detection in the I-DASH project," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, 2011.

[2] A. Maier, J. G. Bauer, F. Burkhardt, and E. Nöth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," *IEEE Acoustics, Speech and Signal Processing*, pp. 1605–1608, 2008.

[3] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoutic and prosodic level information fusion," *Computer Speech and Language*, vol. 27, pp. 151–167, 2012.

[4] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 6, pp. 734–749, 2005.

[5] A. B. Barragáns-Martíneza, E. Costa-Montenegroa, J. C. Burguilloa, M. Rey-Lópezb, F. A. Mikic-Fontea, and A. Peleteiroa, "A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition," *Information Sciences*, vol. 180, pp. 4290–4311, 2010.

[6] S. H. Hsu, M.-H. Wen, H.-C. Lin, C.-C. Lee, and C.-H. Lee, "AIMED - a personalized TV recommendation system," *Lecture Notes in Computer Science*, vol. 4471, pp. 166–174, 2007.

[7] D. Bonnefoy, M. Bouzid, N. Lhuillier, and K. Mercer, ""More Like This" or "Not for Me": Delivering personalised recommendations in multi-user environments," *Lecture Notes in Computer Science*, vol. 4511, pp. 87–96, 2007.

[8] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. Voume 41, pp. 1–24, 2001.

[9] Z.-H. Tan, "Audio and speech processing for data mining," *Encyclopedia of Data Warehousing and Mining - 2nd Edition*, vol. 1, pp. 98–103, 2008.

[10] S. E. Tranter, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1557–1565, 2006.

[11] F. Burkhardt, M. Ekert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," *Proc. 7th International Conference on Language Resources and Evaluation (LREC)*, pp. 1562–1565, 2010.

[12] R. Biuk-Aghai, S. Fong, and S. Yain-Whar, "Design of a recommender system for mobile tourism multimedia selection," *Internet Multimedia Services Architecture and Applications (IMSAA)*, 2008.

[13] G. D., "Genetic and evolutionary algorithms come of age," *Communications of the ACM*, vol. 37, pp. 113–119, 1997.

[14] "Statistics denmark." [Online]. Available: http://www.statistikbanken.dk/statbank5a/default.asp?w=1680

[15] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[16] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 798–807, 2010.

[17] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (version 5.4.32) [computer program]," 2009, available from http://www.praat.org/.

[18] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2095–2103, 2007.

Paper B

Audio-based Age and Gender Identification to Enhance the Recommendation of TV Content

Sven Ewan Shepstone, Zheng-Hua Tan, and Søren Holdt Jensen

The paper was published in the
IEEE Transactions on Consumer Electronics Vol. 59, pp. 721–729, 2013.

The layout has been revised.

Abstract

Recommending TV content to groups of viewers is best carried out when information such as the demographics of the group is made available. However, it can be difficult and time consuming to extract information for every user in the group. This paper shows how an audio analysis of the age and gender of a group of users watching the TV can be used for recommending a sequence of N short TV content items for the group. First, a state of the art audio-based classifier determines the age and gender of each user in an M-user group and creates a group profile. A genetic recommender algorithm then selects for each user in the profile, a single personalized multimedia item for viewing. When the number of items to be presented is different to the number of viewers in the group, i.e. $M \neq N$, a novel adaptation algorithm is proposed that first converts the M-user group profile to an N-slot content profile, thus ensuring that items are proportionally allocated to users, with respect to their demographic categorization. The proposed system is compared to an ideal system where the group demographics are provided explicitly. Results using real speaker utterances show that, in spite of the inaccuracies of state-of-the-art age-and-gender detection systems, the system has a significant ability to predict an item with a matching age and gender category. User studies were conducted where subjects were asked to rate a sequence of advertisements, where half of the advertisements were randomly selected, and the other half were selected using the audio derived-demographics. The recommended advertisements received a significant higher median rating of 7.75, as opposed to 4.25 for the randomly selected advertisements.

1 Introduction

With the merge of DVB-C, DVB-T and DVB-S technologies in recent TV platforms, consumers have become overwhelmed by the sheer amount of content available. A large body of research has therefore looked at various ways of personalizing TV to match the needs of users as far as possible. The Electronic Program Guide (EPG), now an integrated component of most modern television sets, has helped to a small extent to narrow the selection of upcoming programs, and there have been various attempts to personalize the EPG to make it even more effective. However, personalization is not only relevant for the EPG, which is geared towards displaying items that can be selected or scheduled, but also for more dynamic content, such as advertising, trailers and short news clips, which are the glue between program segments on broadcast TV.

In order to be able to recommend content, a user profile is needed. User profiles for recommendation can be extracted explicitly, e.g. through registration questionnaires [1] or by asking users to provide ratings. Data can also be col-

lected implicitly through usage patterns [2], [3], and subsequently fed to a central information server, which can then make recommendations.

Traditional recommender systems then use these profiles, together with metadata and ratings from other users in the network, to provide personalization. One of the issues however, in the context of broadcast TV, is the lack of an uplink channel, through which information such as ratings can be exchanged with the remaining users. It is therefore highly desirable that feedback from users be collected locally, in the set-top box or smart TV if possible, and as unobtrusively as possible, e.g. such as through unobtrusive relevance feedback [3].

By means of local recommendation and implicit user feedback, these systems can work quite effectively, but it is important to consider the preferences of a group of users as well as a single user. This is a particular issue when multiple consumers share a single device, such as a home television, but each has their own user profile and tastes [4]. In the Socially Aware TV Program Recommender for example [5], groups of users who want simultaneous access to the TV are taken into account, where individual profiles that have a common interest are combined.

What is more challenging however, is when multiple viewers share the same TV, but typically only use one person's login, even when a multiple login feature exists, making specification of demographics, extraction of ratings or monitoring for each user difficult to realize in practice. Groups of viewers are further characterized by the fact that users continually come and go, meaning that the TV must quickly adapt itself to the current configuration.

Taking these multiple requirements into account, i.e. local recommendations, implicit gathering of user information and being able to support groups of viewers, one area that has somewhat been overlooked in the context of personalization of multimedia content, is the home audio environment, from which a wealth of user information can be extracted. State-of-the-art feature extraction and modeling techniques, which are in many ways similar to speaker identification systems, make it possible to extract a number of useful attributes from home viewers, from which recommendation profiles can be constructed. In particular, both the age and gender of TV viewers can be extracted.

Determining both the age and gender of speakers is a complicated task and has received considerable attention in recent years. The results achieved are encouraging and are beginning to make it feasible to use this technology as a viable alternative to existing methods of providing user demographics. Age and gender classification systems are generally implemented as a fusion of several subsystems [6], with each subsystem operating using a form of Gaussian mixture model, multilayer perceptrons, hidden Markov models and/or support vector machines [7], [8].

Recent work shows that 3-class gender detection can be done with substantially higher accuracy as high as 75 %, which is roughly 30 % higher than results achieved for 4-class age detection [6]. Here, results for a 7-class classification system also show that separate classes defined as children, young males, young females, adult males, adult females, senior males and senior females can be detected with 61.0

%, 49.4 %, 57.1 %, 27.1 %, 33.8 %, 69.7 % and 53.9 % accuracy, respectively. The largest confusion occurs between young males and adult males, between young females and adult females, between adult females and senior females and finally between children and young females. Even though a lot of room remains for future research to improve these results, there ought to be a substantial basis for recommendation, since the effect of overlapping confusion classes could well be ameliorated by soft market boundaries. For example, in an advertising context, there are many products that would be recommended to both young females and adult females, thus helping to cancel out the confusion overlap seen in these results.

The contribution of this paper is a novel method using on-the-fly detection of the age and gender of the audience present to quickly provide recommendations of TV content to home viewer groups. This is in contrast to other methods that make use of usage data, registration data or questionnaires to obtain the demographics. The focus is on groups of users who are about to be presented with a series of short media items, e.g. between programs. In particular, this work will focus on recommending sequences of advertisements to viewers. The proposed system operates by determining the age and gender class of each user in the group, and then uses this information to find a sequence of content items that best matches the group profile. Ideally, each user should be matched with a content item that belongs to the same age and gender category as that of the user themself. Since the number of advertisements is often predetermined in advance and may not be equal to the number of viewers present, the proposed system ensures that the age and gender demographics will be reflected proportionally in the sequence of advertisements that are about to be presented.

The rest of this paper is organized as follows: The next section introduces the notion of a group TV profile in the context of age and gender demographics, and how existing audio analysis techniques could be combined to construct a group profile. Section 3 introduces a genetic-based recommender, extended to computing of 7-dimensional age and gender ratings, and section 4 demonstrates how an M-user group profile can be adapted to an N-slot advertisement profile, and shows how this is used to drive a genetic algorithm-based recommendation engine. Following this the experimental setting is discussed. The system is then evaluated from a number of perspectives. Finally, conclusions are drawn.

2 Group Profile Derivation

Solving the "Who is sitting in front of the TV?" problem is a challenging task and has yet to be researched fully. When only one person watches TV, attempting to derive additional profile attributes by means of speech or an acoustical analysis does not make much sense, and instead one must rely on other sources of information, such as an explicitly provided user profile, or through image recognition (many households today already have movement detection cameras as a standard games

console accessory).

A typical system could be realized as follows: When multiple users are present, the audio from several microphone pickups in the room is applied to an independent component analysis algorithm that can separate the background TV audio (if any) from the users' speech [9]. Speaker diarization is used on the speech part to separate speaker utterances of different people from one another, and to determine the number of speakers present at any given time [10], [11].

In the ideal case, where the exact age and gender class for each user in the audience is known, a Group Viewer Configuration (GVC) is formed, which can be expressed as follows:

$$GVC = \begin{bmatrix} c_{user_1} \\ c_{user_2} \\ \vdots \\ c_{user_M} \end{bmatrix} \tag{B.1}$$

where c_{user_m} corresponds to the age and gender class of the *m-th* user, $1 \leq c \leq C$, C is the total number of age and gender classes, $1 \leq m \leq M$ and M is the total number of users.

In practice, the speaker utterances from each speaker in the audience are classified according to age and gender to determine their class. However, due to the probabilistic nature in which speaker classification systems work, along with their limited accuracy, it is important to note that each speaker, regardless of their age/gender class, will to some extent be a member of all other classes. In this way, the user profile for a single user m whose real class is c_{user_m}, can be modeled by:

$$\mathbf{x_m} = \begin{bmatrix} p_{m,1} \\ p_{m,2} \\ \vdots \\ p_{m,C} \end{bmatrix} \tag{B.2}$$

where $p_{m,j}$, $0 \leq p_{m,j} \leq 1$, simply represents the actual predicted probability for class j, $1 \leq j \leq C$, and $\sum_{j=1}^{C} p_{m,j} = 1$. The more utterances that can be collected, the better the classification accuracy.

For all M users, a group profile is then constructed from the individual user profiles as follows:

$$\mathbf{X_G} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_M^T \end{bmatrix} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,C} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \cdots & p_{M,C} \end{pmatrix} \tag{B.3}$$

3 Demographic Recommendation

The matching problem can be stated as optimizing the match between the group profile and the sequence of content items (advertisements) that are about to be presented to the users. The basic genetic algorithm approach proposed for extending MacauAp [12] is taken as the starting point for the recommendation system, and performs the relevant matching. Based on user-feedback of categories for tourist destinations, called "spots", the genetic algorithm in MacauAp searches amongst a large number of tourist destinations and finds a sequence with an optimum match between the categories to which the spots belong and the user-liked categories. In the same vein, the purpose of the genetic algorithm for the proposed system is to find the sequence of content items, whose combined demographic profile best matches the audio-derived group profile.

Genetic Algorithms are established computational methods that conduct their searches based on natural selection and genetics, and use the concepts of chromosomes, populations, selection, crossover and mutation [13]. A search is typically initiated by creating a population comprised of units called chromosomes, where each chromosome is effectively a sample of the search space. Each iteration of the algorithm entails selecting two parents from the population (selection) using a tournament or proportionate selection approach. A fitness evaluation is conducted to determine which of the chromosomes ought to be considered as parents. From these parent chromosomes a child chromosome (crossover) is then spawned that comprises attributes from both parents. The child then replaces the weakest chromosome in the population. This process continues until termination, which is usually defined as the point where the population becomes stable. The chromosome with the highest fitness value is then selected as the winner, and is the output of the algorithm. In this approach a chromosome is defined as a sequence of content items (advertisements) that are to be broadcast in the next upcoming break. Each chromosome has N slots, where a slot is defined as a placeholder for a single content item.

Before parents can be selected for crossover, the fitness of each chromosome needs to be computed. As was proposed for the MacauAp scenario, the base fitness for such a chromosome is given as

$$Fitness_{base} = \sum_{i=1}^{N} r_i * Pref_i \tag{B.4}$$

where

$$\begin{aligned} N &= \text{Number of slots in the chromosome} \\ r_i &= \text{Official rating for slot } i \\ Pref_i &= \text{User preference for slot } i \end{aligned}$$

The scalar rating r_i from equation B.4 is extended in this work by converting it to a vectorized form where all age and gender classes are represented:

$$\mathbf{r_i} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_C \end{bmatrix} \tag{B.5}$$

The predefined age and gender ratings for the N content items (equal to the number of slots) is then given as:

$$\mathbf{R} = \begin{bmatrix} r_1^T \\ r_2^T \\ \vdots \\ r_N^T \end{bmatrix} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,C} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N,1} & r_{N,2} & \cdots & r_{N,C} \end{pmatrix} \tag{B.6}$$

For now each user is assumed to be assigned to a single slot, making $M = N$, meaning that each user in the GVC gets to see at least one content item to his/her liking (shortly this will be extended to the case $M \neq N$). Treating $Pref_i = \mathbf{x_m}$, where user m is assigned to slot i, now allows us to express the $Fitness$ more compactly as:

$$Fitness = \mathrm{Tr}(\mathbf{R} * \mathbf{X_G}^T) \tag{B.7}$$

where $\mathrm{Tr}(\mathbf{A})$ is the trace of $\mathbf{A}$, i.e. the sum of the main diagonal of $\mathbf{A}$.

4 Group Profile Adaptation

(B.7) above requires that for each slot, there is a separate set of preferences values. Since however it cannot be assumed that $M = N$ (for example, there might be 4 users present, but 5 advertisements are to be presented), a new set of preferences with the same dimension as N is needed. The intention is to ensure that the each user's demographic membership for all age and gender classes is carried over proportionally to the new preference set. This adaptation process is best explained using the double circle diagram shown in figure B.1, which shows the adaptation process for a single age/gender class[1]. The circle is divided into a fixed number of bins, which run at equally spaced intervals over its entire revolution. On the outer circle one allocates an equal portion of the bins to each user, so for example, for 4 users, there will be 4 separate partitions. The same applies to the inner circle, but instead of allocating bins to users, they are allocated according to slots. For 5 slots one therefore ends up with 5 equally-sized slot partitions. It therefore follows intuitively that for the bins comprising a single slot, rating contributions

[1]The adaptation for each age and gender class is computed independently of the others.

can come from multiple users. The amount that each user contributes to a given slot is directly proportional to the size of the overlap between the user bins and slot bins. Summing over all bins belonging to a single slot partition and dividing by the number of bins per slot, allows one to compute a rating for that slot.

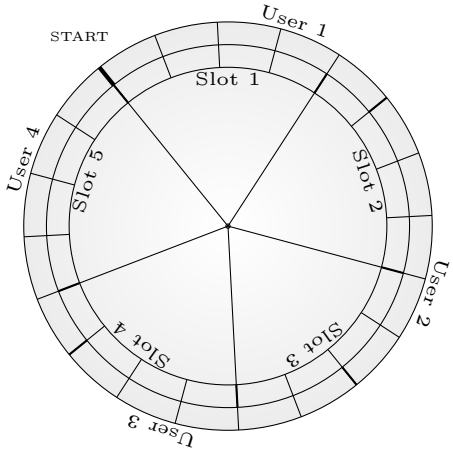


Fig. B.1: Proportional bin selection for a single age-gender class, for 4 users, 5 slots and 20 bins. The rating for a given slot can come from multiple users.

More formally, assuming the number of bins is B, a new rating matrix $\mathbf{\Sigma}_{C,N}$ is defined, where each element $\mathbf{\Sigma}_{i,j}$ is given as:

$$\mathbf{\Sigma}_{i,j} = \frac{1}{\mu} \sum_{k=1}^{\mu} \mathbf{X_G}_{i,\phi(j,k)} \tag{B.8}$$

where

$$B = LCM(M, N), \text{ i.e. the least common multiple} \tag{B.9}$$

$$\mu = \frac{B}{N}, \text{ i.e. slot partition size in bins,} \tag{B.10}$$

$$\nu = \frac{B}{M}, \text{ i.e. user partition size in bins} \tag{B.11}$$

and where

$$\phi(j,k) = \lfloor(((j-1) * \mu + k - 1)/\nu) + 1\rfloor \tag{B.12}$$

represents the user partition to which bin k that is currently being processed, belongs to.

Now armed with separate ratings for each slot, the fitness can now finally be calculated as:

$$Fitness = \sum_{i=1}^{E} \mathbf{R_{i,j}} * \mathbf{\Sigma}_{i,j} \tag{B.13}$$

where

$$
\begin{aligned}
N &= \text{Number of content items (slots) to present} \\
C &= \text{Total Number of age/gender classes} \\
\mathbf{R_i} &= \text{Normalized rating of age/gender class } j \text{ for ad/slot } i \\
\mathbf{\Sigma_{i,j}} &= \text{Normalized group membership for age/gender class } j \text{ for slot } i
\end{aligned}
$$

5 Experimental Setup

5.1 Group Viewer Configurations

To emulate the home group viewers, 50 separate GVCs are defined, which are shown in Table B.1. The viewer configurations that were chosen were based on information provided by Statistics Denmark, which records comprehensive statistics on the composition of Danish households. Here it can be seen that 23.8 % of the population live alone, 38.7 % live with one other person, 14.3 % belong to a family of three, 14.6 % belong to a family of four and 5.5 % belong to a family of five. From these figures, and based on the fact that single-person profiles are excluded, the viewer configurations selected are based on families of two, three and four persons, where the bulk of the distribution lies.

Now just for the two-person households, children and youngsters do not feature much, and only comprise 2.8 % and 2.1 % of households, respectively. In contrast, 37.6 % of households contain adults and 57.5 % have seniors, giving the first ten configurations in Table B.1 below.

Looking at children and youth from just the three- and four-person households, it can be noted that for children, 30.1 % are part of three-person families, but that 69.4 % (more than double) are part of four-person families. For the youth category, 40.1 % of youths belong to three-person families whereas 59.9 % of youths belong to four-person families. Thus it is evident that children and youths should feature fairly strongly in the chosen configurations. The data also shows that there were twice as many seniors with two children living at home (15657 people) than seniors with only one child living at home (7302). Finally, a number of four-person configurations spanning all generations are included. From all this, the two-person configurations 11-17, the three-person configurations 18-30 and the four-person configurations 31-50 in Table B.1 below[2] are constructed.

[2]Note that one differentiates between the number of people in the household, and the number of viewers in front of the TV, e.g. there will be multiple 2-person and 3-person configurations for a 4-person household.

Profile No	1	2	3	4	5
Profile	AM+AM	AM+AF	AF+AF	SM+SM	SM+SF
Profile No	6	7	8	9	10
Profile	SF+SF	AM+SM	AM+SF	AF+SM	SF+SF
Profile No	11	12	13	14	15
Profile	C+C	C+YM	C+YF	C+AM	C+AF
Profile No	16	17	18	19	20
Profile	C+SM	C+SF	C+C+SM	C+C+SF	C+SM+SF
Profile No	21	22	23	24	25
Profile	C+C+YM	C+C+YF	C+C+AM	C+C+AF	C+YM+YM
Profile No	26	27	28	29	30
Profile	C+YF+YF	C+YM+YF	C+AM+AM	C+AM+AF	C+AF+AF
Profile No	31	32	33	34	35
Profile	C+C+C+YM	C+C+C+YF	C+C+C+AM	C+C+C+AF	C+C+C+SM
Profile No	36	37	38	39	40
Profile	C+C+C+SF	C+C+C+C	C+C+YM+YM	C+C+YM+YF	C+C+YM+AM
Profile No	41	42	43	44	45
Profile	C+C+YF+AF	C+C+AM+AF	YM+YM+AM+AF	C+YM+AM+AF	C+AM+AF+SM
Profile No	46	47	48	49	50
Profile	C+AM+AF+SF	C+C+SM+SF	YM+YM+SM+SF	YM+YF+SM+SF	AM+AF+SM+SF

Table B.1: Selected TV viewer configuration. C=Child, YM=Young Male, YF=Young Female, AM=Adult Male, AF=Adult Female, SM=Senior Male, SF=Senior Female

5.2 Audio Classification of Age and Gender

For each speaker from the viewer configuration profile, a set of real speaker utterances are classified to determine their class. The utterances are selected by randomly picking out a speaker with a matching class from the evaluation portion of the aGender corpus [14]. The aGender corpus was supplied to participants in the InterSpeech 2010 Paralinguistic Challenge to enhance the development of age and gender algorithms. The training part of the dataset contains 32527 utterances from 472 speakers, the development part contains 20549 utterances from 300 speakers and the testing part contains 17332 utterances. It comprises 4 age classes: children (7-14 years), young people (15-24 years), adults (25-54 years) and seniors ($\geq$ 55 years), and 3 gender classes: children, males and females. Children are classed as their own gender since males are indistinguishable from females at that age. In more recent work, the age boundaries are slightly different, i.e. children ($\leq$ 13 years), young people (14-19 years), adults (20-54 years) and seniors ($\geq$ 55 years) [6]. The latter age boundaries, corresponding to the recent work, were chosen[3].

For the speaker that was selected, the speaker utterances for the speaker were pooled to form a contiguous segment. Each speech segment was then submitted for classification, to determine its class. The speaker results were then combined to form the group profile $\mathbf{X_G}$ from above.

The audio classification system is constructed as a hybrid system comprising two subsystems. The first subsystem models acoustic speaker features and the second subsystem models the prosodic features. Modeling several feature types

[3]The original aGender age boundaries were chosen solely on the basis of marketing aspects, and not on any physiological aspects.

increases the classification accuracy of the system.

The acoustic subsystem is modeled using the well-known Gaussian Mixture Model Universal Background Model (GMM-UBM) approach [15]. After voice activity detection was applied to each utterance [16], feature extraction was performed using 13-dimensional Mel Frequency Cepstral Coefficients (including C0), with 1st and 2nd derivative, to give 39 coefficients per 25 ms frame (15 ms overlap). Mel Frequency Cepstral Coefficients (MFCCs) are simply a compact representation of the spectral envelope of a speech signal. A 512- component GMM-UBM was trained using all the training data from the aGender corpus. Seven speaker models, one for each class, were then adapted from the single UBM using the training data from each class. For the adaptation process, a relevance ratio of 12 was used. The accuracy for the acoustic subsystem for all classes was 49.9 %.

To model the prosody features the prosody baseline referred to as System 7 in a previous work was used [6], which models prosody features at the syllable level instead of at the frame level. The syllable boundaries are determined as follows: For each utterance, all frames are marked as voiced or unvoiced (unvoiced where the pitch is undefined) and all unvoiced frames are discarded. For the remaining frames, the normalized energy contour is used as a key to determining the syllable boundaries, where valleys in the contour indicate the start of a new syllable.

The prosody features modeled for each syllable are contours of pitch, energy, formants, syllable duration and spectral harmonic energy (obtained from the power spectrum at harmonics of F0). The Praat package was utilized to extract pitch and energy features from each utterance and Matlab was used to compute the spectral harmonic energy. After applying time scale normalization for the interval -1 to 1, the contours were then modeled as sixth-order Legendre polynomials, meaning that instead of an entire contour, only six coefficients need to be stored [17]. Seven 512-component GMM models were then trained with the prosody features, one for each class. The prosodic subsystem's accuracy for all classes was 42.0 %.

For the hybrid system, the acoustic and prosodic subsystems were combined using weighted summation-based fusion [6] of the subsystem results. The hybrid classifier model was tested on the entire development data set, an accuracy was achieved on the combined system of 50.0 %. As a comparison, another work using seven individual subsystems was able to attain an accuracy of 50.3 % [6][4].

5.3 Initial Rating of Ads

The advertisement corpus used in this paper was provided courtesy of TV2, a nationwide television broadcaster in Denmark. The commercials are subdivided into 24 categories. Examples of categories are *Food, Beverages*, *House and Home*

[4]Whereas the hybrid system appears to only offer a marginal increase in accuracy over the acoustic-only system, it should be borne in mind that they represent the average for *all* 7 age and gender classes for each system, and that the response for individual classes for the two systems are in some cases quite different.

and *Media*.

A random subset of advertisements was taken from each category to give a total of 200 commercials. These were then split into four separate groups. For each group of 50 advertisements, three subjects were asked to rate 50 of them, on the basis of how well they matched the seven age and gender categories listed above. The scale used was the standard 1-5 likert scale, with 1 being not-relevant and 5 most relevant. The ratings were then averaged for each advertisement across the participants for each group, by taking the median.

Table B.2 shows a sample selection of the rated ads. Commercial details of advertisements have been withheld.

Advertisement	C	YM	YF	AM	AF	SM	SF
Washing Machine	1	3	3	5	2	4	1
Computer Game	4	5	3	4	1	2	1
Alcoholic Drink	1	3	5	1	4	3	2
Lift Chair for Elderly	1	1	1	1	1	4	5
Children's Building Bricks	4	4	2	3	2	1	1
Ferry Company	2	2	1	5	5	5	4
Retail Bank	1	2	3	5	4	1	1

Table B.2: Selected commercials along with their ratings. C=Child, YM=Young Male, YF=Young Female, AM=Adult Male, AF=Adult Female, SM=Senior Male, SF=Senior Female

5.4 Recommendation of Ads

The matching of advertisements was carried out as follows: The advertisements were combined from the four rating groups to give a total of 200 advertisements. The genetic algorithm was initialized with 50 chromosomes, with each chromosome comprising N randomly chosen advertisements. It was not possible for the same advertisement to appear twice within a given chromosome. After this, 500 iterations of genetic selection were run, where the fitness in each round was recomputed according to the predefined advertisement's age and gender ratings and the extracted group profile. At the end of the entire run, the chromosome with the highest fitness was selected as the winner and used as the sequence of recommended advertisements.

6 Evaluation and Results

6.1 Evaluation of User Categories

In the first part of the evaluation, a system using an audio-derived group profile (the proposed system) is compared to an ideal system, where the group profile

matching a given GVC is provided explicitly, for example through an online user form or registration questionnaire.

In the proposed system an audio classifier determines the M-user group profile by connecting real speaker utterances to each user in the GVC and computing a probabilistic membership for each class. In the ideal system, there is no audio classification, and instead each user profile $\mathbf{x_m}$ is constructed by setting a value of 1 for the class matching the user category, and a 0 for the remaining classes. The users are combined in the same way to form the group profile.

The objective for either system is for a given GVC, to recommend a sequence of content item where for each user, a single matching advertisement is suggested that has the same age and gender category. Since in this case each user maps to a single ad, $M = N$ and no group profile adaptation is carried out in this particular evaluation. The ultimate category for an advertisement is simply defined as the advertisement's age and gender class that received the highest rating. Some of the advertisements, however, had two or more classes that had received the same, highest rating. In these cases the advertisements were assigned to multiple categories.

The overall effectiveness of the system in predicting the correct category was measured as follows: for each advertisement that was recommended, an input-output pair was formed containing the true user category, e.g. *Child*, and the category of the advertisement that was recommended, e.g. *Young female*. To take account of the advertisements that had multiple categories, additional input-output pairs were created as necessary. For each advertisement to user recommendation, there is therefore at least one mapping from input class to output category. For each GVC, 50 evaluations were carried out, with different speaker utterances being used in each evaluation (with possible replacement).

Once all 50 GVCs had been tested, the input-output pairs were collated and entered into a 7x7 contingency table. To measure the level of association between the input categories (the rows, R) and the output categories (the columns, C), Pearson's chi-squared test was carried out. The strength of association using Cramer's Phi (or V) for categorical variables can be computed for each system as:

$$\phi_c = \sqrt{\frac{X^2}{T(k-1)}} \tag{B.14}$$

where

$$\begin{aligned} X^2 &= \text{Pearson's chi-squared statistic} \\ T &= \text{Total number of input-output pairs} \\ k &= min(R, C) \end{aligned}$$

In the experiments, T is computed as:

$$T = Sim_{count} * GVC_{count} * N_{gvc} * Cat_{count} \tag{B.15}$$

where

$$Sim_{count} = \text{Number of separate group simulations}$$
$$GVC_{count} = \text{Number of GVCs}$$
$$N_{gvc} = \text{Number of content items recommended}$$
$$Cat_{count} = \text{Number of highest rated categories for each item}$$

Furthermore the value of k is fixed at 6, since $R = C = 7$. The obtained values for each system are shown in Tables B.3 and B.4 below.

	C	YM	YF	AM	AF	SM	SF	**4595.0**
C	3175.0	747.0	626.0	19.0	25.0	1.0	2.0	**4595.0**
YM	288.0	761.0	553.0	231.0	152.0	32.0	34.0	**2051.0**
YF	61.0	236.0	446.0	96.0	255.0	19.0	57.0	**1170.0**
AM	1.0	69.0	59.0	899.0	480.0	395.0	185.0	**2088.0**
AF	0.0	39.0	135.0	399.0	850.0	146.0	315.0	**1884.0**
SM	0.0	20.0	20.0	519.0	239.0	698.0	388.0	**1884.0**
SF	0.0	16.0	55.0	280.0	550.0	406.0	749.0	**2056.0**
Totals	**3525.0**	**1888.0**	**1894.0**	**2443.0**	**2551.0**	**1697.0**	**1730.0**	**15728.0**

Table B.3: Contingency Table for Case 1: Upper Bound System - Explicitly Provided age and Gender Profile. $X^2 = 18293$. $\phi_c = 0.4403$. $T = 15728$.

	C	YM	YF	AM	AF	SM	SF	**4595.0**
C	2585.0	1452.0	1698.0	367.0	559.0	115.0	183.0	**6959.0**
YM	98.0	537.0	440.0	587.0	442.0	284.0	246.0	**2634.0**
YF	186.0	234.0	381.0	118.0	247.0	45.0	96.0	**1307.0**
AM	28.0	344.0	291.0	841.0	558.0	585.0	374.0	**3021.0**
AF	87.0	210.0	422.0	396.0	755.0	276.0	495.0	**2641.0**
SM	16.0	171.0	157.0	661.0	439.0	614.0	400.0	**2458.0**
SF	60.0	152.0	261.0	443.0	667.0	407.0	583.0	**2573.0**
Totals	**3060.0**	**3100.0**	**3650.0**	**3413.0**	**3667.0**	**2326.0**	**2377.0**	**21593.0**

Table B.4: Contingency Table for Scheme 2: Proposed System - Audio-derived Age and Gender Group Profile. $X^2 = 9295$. $\phi_c = 0.2678$. $T = 21593$.

The results show that the effect size $\phi_c = 0.2678$ of the audio-based system is lower than that of the theoretical maximum given by the ideal system of $\phi_c = 0.4403$, which can be accounted by the error introduced by the audio classifier. Depending on what speaker utterance was classified, there will always be varying degrees to which a given class is a member of the other classes (never 0 or 100 %). This can also clearly be seen in the diagonals of each table, which represent the number of hits correctly classified for each user category. In the ideal system, there are (as expected) a larger number of correctly classified advertisements.

What is interesting to note for the proposed system, however, is that there is a significant correlation between the input categories and output categories. This is in spite of the accuracy only being half that of the theoretical maximum. To test for significance, the null hypothesis may be stated that the age and gender categories

of the ads for each user are chosen with equal probability. With a goodness of fit of $X^2 = 9295$ and $df = 36$, it was found that with a $p < 0.1$ that the null hypothesis is disproved. Furthermore, the value of $\phi_c = 0.2678$ for the proposed system is over 2.5, which according to the threshold values for Cramer's V, corresponds to a very strong association between the true categories and the recommended categories.

What is also interesting to note in the proposed system is the value of T, which happens to be 27 % higher than in the ideal system, meaning more multi-category advertisements were selected in the proposed system. The reason for this is believed to be the classifier-induced increased membership of each class of every other class, which leads to selecting advertisements with multiple highly-rated classes.

6.2 Testing the Effectiveness of the Group Adaptation Approach

When group-to-slot adaptation is performed, $M \neq N$ and the number of advertisements to be recommended is different to the numbers of users sitting in front of the TV. Since the number of advertisements for a given ad break are predetermined, a system that does not employ the adaptation technique would have to resort to other method to fill up the additional slots. In this part of the evaluation two systems are compared: a system where the remaining $N - M$ slots are filled with random advertisements, and a system where the full group profile adaptation technique is applied.

To formally evaluate to what extent each user's age and gender class is represented in the sequence of recommended items, and to determine whether group profile adaptation gives a better proportional representation than the alternative system, an adapted version of the group profile $\mathbf{X_G}$ is correlated with the age and gender ratings of the recommended advertisements. The idea behind this is that the more accurately the individual user's classes are reflected in the sequence of items that are presented, the stronger the correlation will be. For example, if the initial GVC was $[C, C, C, SF]$, meaning that three quarters of the audience are children, then it is expected that three quarters of the advertisements will also be targeted to children.

To allow for this comparison, each GVC is converted to a 7-dimensional age- and gender representation x_{gvc}, where a 1 is given for each age-and-gender class in the GVC, and a 0 is given otherwise. In doing so, the class that has the strongest representation in the GVC will end up having the largest weighting. Likewise a similar 7-dimensional age and gender representation x_{items} is constructed for the N advertisements that are represented. To determine the weighting for each class, each advertisement's ratings for all classes are summed across each individual class. In this way, the class that has the strongest representation across all N items will end up having the largest weighting.

To compute the correlation between between x_{gvc} and x_{items}, Kendall's Rank

Correlation Coefficient is used[5], which is a non-parametric hypothesis test that measures the degree of concordance between the values being compared. The test ensures that before the strength of the correlation is computed, the data on both sides is ranked, and where tie ranks are observed, the rank value simply becomes the average of the individual ranks. For each τ_B value that is computed, the corresponding z-score is also computed, which is characterized by a normal distribution when the variables are statistically independent.

The results show that there is a stronger overall effect, and hence preservation of the original age and gender classes, when group profile adaptation is applied ($\tau_B = 0.2316, Z_B = 34.29$) than when random advertisements are used to fill the remaining slots ($\tau_B = 0.08, Z_B = 11.91$). The values for τ_B and Z_B were also calculated for each of the 50 GVCs for both schemes. When the direction of correlation is taken into account (zero correlation is considered better than a negative correlation), it was noted that in 47 out of 50 cases that a stronger rank correlation coefficient (and accompanying z-score) was obtained for the case where group profile adaptation was employed. Therefore, from a statistical standpoint, there is a stronger overall effect introduced when applying adaptation. Table B.5 shows the values for the first 10 individual τ_B and Z_B values.

Profile	$\tau_B(1)$	$\tau_B(2)$	$Z_B(1)$	$Z_B(2)$	Profile	$\tau_B(1)$	$\tau_B(2)$	$Z_B(1)$	$Z_B(2)$
1	0.4436	0.4684	8.948***	9.447***	2	0.5194	0.5509	10.49***	11.12***
3	0.4489	0.4418	9.061***	8.935***	4	0.1069	0.3847	2.157	7.756***
5	0.133	0.2334	2.684	4.708***	6	0.116	0.3087	2.339	6.238***
7	0.4067	0.5548	8.213***	11.2***	8	0.2525	0.3182	5.095***	6.417***
9	0.2914	0.314	5.88***	6.34***	10	0.1117	0.2727	2.254	5.509***
11	-0.1118	0.2866	-2.247	5.797***	12	-0.2833	0.2034	-5.7	4.095***
13	-0.1052	0.4062	-2.119	8.21***	14	-0.03915	0.1746	-0.7879	3.508**
15	0.04066	0.2132	0.8205	4.297***	16	-0.3315	-0.2701	-6.674	-5.421
17	-0.2746	-0.1601	-5.529	-3.221	18	-0.3157	-0.1439	-6.555	-2.99
19	-0.2225	-0.0544	-4.626	-1.13	20	-0.3726	-0.4201	-7.498	-8.431
21	-0.1419	0.2989	-2.952	6.227***	22	0.1309	0.4681	2.721*	9.78***
23	0.01513	0.1486	0.3143	3.096***	24	0.1175	0.2837	2.443	5.913***
25	-0.1934	0.0474	-4.025	0.9879	26	0.1512	0.4061	3.145**	8.481***
27	-0.02529	0.3764	-0.5089	7.585***	28	0.02998	0.09813	0.6237	2.04
29	0.2582	0.2149	5.195***	4.317***	30	0.148	0.175	3.084*	3.647**
31	0.1604	0.3741	3.33***	7.807***	32	0.3073	0.5142	6.399***	10.74***
33	0.1227	0.2244	2.547	4.676***	34	0.3138	0.3641	6.52***	7.593***
35	-0.2107	-0.01892	-4.378	-0.3941	36	-0.121	0.01859	-2.512	0.3868
37	0.2442	0.4158	4.924***	8.406***	38	-0.07189	0.2175	-1.445	4.383***
39	0.1548	0.3932	3.255**	8.265***	40	0.123	0.1834	2.576*	3.846***
41	0.3394	0.445	7.127***	9.36***	42	0.1567	0.2059	3.286**	4.323***
43	0.4174	0.4247	8.772***	8.93***	44	0.1968	0.291	3.958***	5.859***
45	0.1803	0.1432	3.634**	2.887**	46	0.1206	0.1621	2.427	3.265**
47	-0.4925	-0.3729	-10.31	-7.803	48	-0.08412	-0.04439	-1.768	-0.9326
49	-0.1903	-0.168	-3.835	-3.381	50	0.566	0.6167	11.42***	12.43***

Table B.5: Values of τ_B and Z obtained for both schemes, shown for all 50 group viewer configurations. One-tailed p-value indications are *: $p < 0.005$; **: $p < 0.001$; ***: $p < 0.0001$

[5]Kendall's τ_B is considered a better statistic for smaller amounts of data than Spearman's Rank Coefficient, and where the ranks that are to be compared have ties.

6.3 User Study Evaluation

In another evaluation relating to this study [18], a user study was conducted where subjects were given the chance to evaluate the proposed system. Due to time constraints, it was not possible to evaluate all 50 group configurations, and therefore a subset of 11 configurations was selected for the study. A total of 12 subjects were asked to participate in the study.

Subjects were shown several different GVCs, and for each GVC, a sequence of 10 advertisements. Five of the advertisements were recommended using a system where the hybrid audio classifier, group profile adaptation algorithm and genetic selection algorithm were present in the system. The other five advertisements were randomly selected, without replacement. Subjects were however, not informed which advertisements were recommended and which were randomly selected. They were then asked to rate on a scale of 1-10 (1 completely irrelevant and 10 most relevant) on whether they thought a given advertisement was suitable for any of the the members of the GVC. A 10-point scale was used to ensure that subjects took a non-neutral stance when rating. For example if the subject thought the advertisement appealed highly to children, and the child category was part of the GVC, then naturally the advertisement would receive a higher rating.

The results show that on average the recommended ads received a higher median rating of 7.75 than the randomly selected ones, which received a rating of 4.25. To test the statistical significance of the items receiving higher ratings, let x represent all samples corresponding to the median ratings of all users for the randomly-selected items and let y represent samples corresponding to the median ratings of all users for recommended items, and test the null hypotheses that $y - x$ comes from a distribution of zero median. Treating the rating scales as ordinal, the 2-sided Wilcoxen Signed Rank test is used to test for significance. With a z-score $z = -2.628$ and $p < .01$ there is a significant increase in the median rating for each test group, thus disproving the null hypothesis. Finally, the effect size using Pearson's correlation coefficient $r = \frac{Z}{\sqrt{N}}$ is also computed, where Z is the z-score from above and $N = 24$ is the number of observations, and is found to be $r = -0.535$. Since the absolute value is above Cohen's benchmark of 0.5, it can be concluded that using the age-and-gender analysis approach has a large effect on the user ratings.

7 Conclusion and Future Work

This paper showed how an audio classifier can be used to elicit a demographic group profile from a given audience, and how this can be used to provide recommendations. Even at the level of state-of-the-art age and gender detection, which is about 50 %, there is good potential in using audio analysis for recommendation. For the proposed system, it was found that there was a strong relationship between the true user categories and the recommended advertisement categories.

In the majority of cases, group profile adaptation leads to a stronger reflection of the users' age and gender classes than simply adding random advertisements to the remaining slots. User studies confirm that the strength of the recommendation can be perceived and that the recommended advertisements were more suitable than randomly selected advertisements.

Finally, it is proposed that the system be used as a baseline for future work. This includes an investigation into further novel ways in which the accuracy of detecting the age and gender of viewers can be enhanced, with the intention to see to what extent it is possible to approach the upper bound results for the explicitly-provided group profile system.

Acknowledgment

The authors wish to thank TV2 Denmark for providing the video TV commercials and Felix Burkhardt for supplying the aGender corpus.

References

[1] S. H. Hsu, M.-H. Wen, H.-C. Lin, C.-C. Lee, and C.-H. Lee, "AIMED - a personalized TV recommendation system," *Lecture Notes in Computer Science*, vol. 4471, pp. 166–174, 2007.

[2] Y.-C. Chen, H.-C. Huang, and Y.-M. Huang, "Community-based program recommendation for the next generation electronic program guide," *IEEE Transactions on Consumer Electronics*, vol. 55, pp. 707–712, 2009.

[3] M. Z. Bjelica, "Unobtrusive relevance feedback for personalized TV program guides," *IEEE Transactions on Consumer Electronics*, vol. 57, pp. 658–663, 2011.

[4] D. Bonnefoy, M. Bouzid, N. Lhuillier, and K. Mercer, ""More Like This" or "Not for Me": Delivering personalised recommendations in multi-user environments," *Lecture Notes in Computer Science*, vol. 4511, pp. 87–96, 2007.

[5] C. Shin and W. Woo, "Socially aware TV program recommender for multiple viewers," *IEEE Transactions on Consumer Electronics*, vol. 55, pp. 927–932, 2009.

[6] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoutic and prosodic level information fusion," *Computer Speech and Language*, vol. 27, pp. 151–167, 2012.

[7] T. I. Meinedo H., "Age and gender detection in the I-DASH project," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, 2011.

[8] A. Maier, J. G. Bauer, F. Burkhardt, and E. Nöth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," *IEEE Acoustics, Speech and Signal Processing*, pp. 1605–1608, 2008.

[9] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. Voume 41, pp. 1–24, 2001.

[10] Z.-H. Tan, "Audio and speech processing for data mining," *Encyclopedia of Data Warehousing and Mining - 2nd Edition*, vol. 1, pp. 98–103, 2008.

[11] S. E. Tranter, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1557–1565, 2006.

[12] R. Biuk-Aghai, S. Fong, and S. Yain-Whar, "Design of a recommender system for mobile tourism multimedia selection," *Internet Multimedia Services Architecture and Applications (IMSAA)*, 2008.

[13] G. D., "Genetic and evolutionary algorithms come of age," *Communications of the ACM*, vol. 37, pp. 113–119, 1997.

[14] F. Burkhardt, M. Ekert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," *Proc. 7th International Conference on Language Resources and Evaluation (LREC)*, pp. 1562–1565, 2010.

[15] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[16] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 798–807, 2010.

[17] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2095–2103, 2007.

[18] S. E. Shepstone, Z.-H. Tan, and S. H. Jensen, "Demographic recommendation by means of group profile elicitation using speaker age and gender detection," *Proceedings of Interspeech, Lyon, France*, pp. 2027–2031, 2013.

Paper C

Using Audio-derived Affective Offset to Enhance TV Recommendation

Sven Ewan Shepstone, Zheng-Hua Tan, and Søren Holdt Jensen

The paper was published in the
IEEE Transactions on Multimedia Vol. 16, pp. 1999–2010, 2014.

The layout has been revised.

Abstract

This paper introduces the concept of affective offset, which is the difference between a user's perceived affective state and the affective annotation of the content they wish to see. We show how this affective offset can be used within a framework for providing recommendations for TV programs. First a user's mood profile is determined using 12-class audio-based emotion classification. An initial TV content item is then displayed to the user based on the extracted mood profile. The user has the option to either accept the recommendation, or to critique the item once or several times, by navigating the emotion space to request an alternative match. The final match is then compared to the initial match, in terms of the difference in the items' affective parameterization. This offset is then utilized in future recommendation sessions. The system was evaluated by eliciting three different moods in 22 separate users and examining the influence of applying affective offset to the users' sessions. Results show in the case when affective offset was applied, that better user satisfaction was achieved, where the average ratings went from 7.80 up to 8.65, with an average decrease in the number of critiquing cycles which went from 29.53 down to 14.39.

1 Introduction

Even with the steady increase of on-demand services such as Netflix and HBO[1], broadcast TV is still firmly entrenched in the home. It is typically the place where the local news and programming is to be found, where many consumers would be reluctant to part with. It is easy to use - turn on the TV, find a channel and watch. Since the consumer does not take part in the selection of the program lineup, recommendations can be serendipitous, something that customers value. From the provider-side there have been substantial investments in satellite, terrestrial and cable networks, and they want to see the best return on investment. Thus it is anticipated that broadcast TV will not be going away any time soon.

The Electronic Program Guide (EPG) is today an integrated part of most home television sets and set top boxes, and is still the predominant method when it comes to navigating both currently-showing and up-and-coming TV programs in the broadcast realm. It is typically presented in a grid-like fashion, with channels down and programs across the grid. However, while the EPG does provide the consumer some assistance, there can still be an overwhelming amount of content to choose from. Not only is the currently-airing program of interest, but also future programs that the consumer may wish to record or be reminded about. To

[1] Both Netflix and HBO offer broadband delivery over IP networks (HBO is traditionally a cable and satellite TV provider, but also offers broadband delivery through their on Demand service).

illustrate: over a three-hour period with 30 available channels, with a program length of 30 minutes (typical during prime-time viewing), there are 180 programs to choose from, making an informed decision difficult.

In order to recommend something personal, a user profile is needed. The user profile data can be collected explicitly, e.g. by requesting users to supply data, or implicitly, through usage patterns. Matching of the user profile to the potentially recommendable content of interest can take place at two levels. At the cognitive level, semantic information such as content descriptors, e.g. genre or user ratings are utilized. The affective level on the other hand deals with the emotional context of the user, and how this relates to the content. The notion of cognitive and affective levels is not a new idea, and has been proposed before in the context of video content retrieval [1].

One area that has received little attention, in the context of recommending content within the EPG framework, is using the user's direct audio environment to extract profile information that can be used to make recommendations. State-of-the-art speaker recognition methods have made it substantially more feasible to extract information about the users, such as their age and gender [2] or emotions [3], using models built upon a text-independent speaker recognition framework.

In this paper we propose a novel framework that takes into account users' audio derived moods to provide the most relevant TV channel recommendation to them. A state-of-the-art audio classifier classifies users' speech into individual emotions, which contribute ultimately to their mood. Since, for a given mood, two separate users might have different ideas of what would be applicable to watch, we do not expect them to find the initially recommended item immediately appealing [4]. Users are therefore given the possibility to critique the item by navigating the emotion space of all candidate items to find a more suitable item, should they wish to do so. To quantify the difference between the initial item and finally selected item, we model what we call the *affective offset* between the items. The novelty lies in leveraging this affective offset to provide system adjustments in such a way that future recommendations are more tailored to the individual person.

This paper is organized as follows: Section 2 starts with a discussion on psychological emotion theory, and how this relates to the proposed framework. Section 3 gives an overview of emotion detection in speech. The following section then introduces critique-based recommender systems. Section 5 presents the recommendation framework, discussing aspects relating to mood detection, critiquing and affective offset. Section 6 presents our experimental work and the following section discusses the findings. The final section concludes the paper, and provides some recommendations for future work.

2 Moods and Emotions

Since moods cannot be measured in the same way as emotions, there is still a lot that is not yet understood about moods. We know however, that emotions, which are more transient in nature, give rise to moods and that certain emotions, such as anger, can cause one to be in a bad mood for a longer period. If certain emotions are experienced strongly and often enough over a given time period, they might eventually give rise to moods, e.g. a continuous sequence of events that cause irritation might lead one to be in a bad mood. A person with a propensity for being in bad moods, might more easily be triggered into becoming angry. While there is no agreement in the literature on how long a mood lasts, it is generally understood that moods last longer than emotions [5].

There is a difference between the mood a person is in and the *pervasive mood* of the content item they might want to see [4]. Mood management theory suggests that people will generally follow their hedonistic desires [6], meaning for example, that somebody in a bad mood might want to watch good mood content to *repair* their negative affective state.

Since we cannot measure mood directly, we concern ourselves with the actual emotions, and how they might be used to determine an entry mood for the system. A person's emotional state can be acquired either explicitly or implicitly. Due to problems seen with explicit acquisition of emotions [4], it has been suggested that they be collected implicitly. Of all the induction methodologies available for obtaining the emotional state, speech is the cheapest and most non-intrusive method.

While the modeling of emotions themselves has always been a very controversial topic [7], the most prominent model used is the dimensional approach, which is based on separate *valence*, *arousal* and *dominance* dimensions, where any emotional state can be represented as a linear combination of these three basic dimensions. Recent studies show that the valence and arousal axes account for most of the variance [1] and that these are typically the two prominent dimensions used in digital systems. We follow in the same vein. In the *VA* space, valence is more commonly referred to as the pleasantness of the emotion, whereas arousal refers to the actual intensity.

The well-known dimensional model known as the Circumplex Model of Affect [8] which is also based on valence and arousal shows how emotional states exhibit a very particular ordering around the periphery of a circle. Emotions that are close in nature are adjacent to one another, whereas bipolar emotions are situated on opposites sides of the circle. Furthermore the emotional states are not distributed evenly around the circle, and some states lie closer to each other on the circumplex than others. The location of these affective states has been determined using empirical data from psychological studies. Each location is expressed in degrees going counterclockwise around the circle, starting at 0° from the positive valence

axis. While there is general agreement on the location of the emotional states, several studies have concluded different exact locations, and recent updates to these models have been made using more stringent statistical models [9].

Not only are there different interpretations of the locations of these states, but very interestingly, the very orientation of the valence-arousal axes has been debated [10]. Some studies have proposed shifting the axes, for example, by orienting them at a 45° angle, or by placing the axes where the emotions are most densely clustered.

While the valence-arousal model is well suited to Human Computer Interaction (HCI) applications, distinct emotion categories, as used in most emotion speech databases today, are not. It can therefore be difficult to relate these fixed categories to the valence arousal *VA* space. Furthermore, labeling of elicited emotions with universal labels has come under scrutiny [7], where it has been postulated that the actual felt emotions, for example, as shown in physiological readings, such as increased heart rate, might not be the same as the emotion labels themselves. This has especially been demonstrated with studies from non-western cultures. A previous work has for example looked at mapping from the *VA* space to distinct emotion categories, using clustering with density estimation [11]. However not only is this more in the context of affective video labeling, but it relies on an intuitive interpretation of what emotion each cluster is assigned to. This can be particular tricky for emotions very close to one other, and where the ordering of the clusters might change, such as in the case for the emotions fear and anger.

This study uses the Circumplex Model of Affect to model the fixed emotions and the *VA* space to model the content items. The Circumplex Model of Affect has the advantage of treating emotion categories as single points around a circle while at the same time giving sense of location, and ordering, for the emotions. Furthermore, since emotions points are relative to the valence arousal axes, the model gives an easy interpretation of what happens when the valence arousal axes are shifted, or tilted. All this will help us to relate the emotion categories to the *VA* space shortly.

3 Detecting Emotions in Speech

Emotion classification in speech is a challenging task and has received a lot of attention in the past ten years. While there is recent interest in continual modeling of emotions [12], speech utterances are generally assigned to fixed labels, such as Ekman's "big six" emotions (anger, disgust, fear, happiness, sadness and surprise), and emotion speech datasets (corpora) typically contain either acted speech [13] [14] or spontaneous speech [15] assigned to fixed emotion labels.

After any necessary speech-signal pre-processing, low-level feature descriptors are extracted, from which an appropriate model can be constructed. Many parameters are used to detect emotion, including mel-frequency cepstral coefficients (MFCCs), which have been the most investigated features for emotion recognition.

MFCCs are simply a compact representation of the spectral envelope of a speech signal. In addition to MFCCs, pitch, intensity, formants and even zero-crossing rate are used. Furthermore the modeling can either be based on fixed length features or variable length features.

Emotions are modeled using a wide variety of techniques including Gaussian mixture models (GMMs), support vector machines and back propagation artificial neural networks. Two recent methods for modeling emotions include class-specific multiple classifiers, based on standard modeling techniques [16], and modeling of emotions using front-end factor analysis (i-vectors) [3] [17].

In the i-vector model, each utterance is expressed as a low-dimensional i-vector (usually between 10 and 300 dimensions). One of the advantages of modeling in the i-vector space is that the i-vectors themselves are generated as unsupervised data [18], without any knowledge of classes. What this essentially means is that when emotion classes are enrolled, a more traditional classifier, such as a Support Vector Machine (SVM) can be used, allowing for quick enrollment of the users' emotional data. This can be an advantage when lots of background data is needed to increase the classification performance. In the i-vector-based system, the background data can be incorporated in the training of the GMM and total variability model, which are used to extract the i-vectors themselves, and which then need not be retrained. Potentially this can reduce modeling of the emotion classifier from hours to seconds. In this work, we have elected to use the i-vector model for emotion classification.

4 Critique-based Recommender Systems

Since typically the content from only one channel can be consumed at any given point, there is a strong basis for providing recommendation for EPG items by quickly being able to select the most relevant channel.

There have for example been works that have looked at recommending content within the EPG framework, that rely both on collaborative [19] as well as content-based [20] techniques. In particular, collaborative recommender systems rely on using other people's ratings for content to generate a list of recommendations for a user. However, we do not believe these fit in well within the EPG framework. Firstly, there is an out-of-band (with the broadcast stream) exchange of ratings between users that needs to take place. While this may seem trivial with today's permanently connected TVs, it is an overdimensioned solution. Secondly, and most importantly, the very nature of broadcast TV is that much of the content that is broadcast may be short-lived and it is possible that it will never be rebroadcast. Once the program has aired, there would be little interest in other users ratings for the program, had these been collected in the first place.

Knowledge-based recommender systems came into existence to deal with the problem of how to recommend knowledge intensive items such as customizable digital cameras, for which ratings might not be easy to acquire, or where they might

not be entirely applicable for the given application [21]. An inherent assumption with knowledge-based systems is that a user may be somewhat undecided on what to search for, and it is therefore the task of the system to guide the user to the item of interest. In a typical case-based recommender, a form of knowledge-based system, a process known as *critiquing* is used in the following manner:

1. The consumer's preference is extracted, either explicitly, or implicitly.
2. Using some sort of similarity metric, the system provides an initial recommendation.
3. The consumer either accepts the recommendation, which ends the entire process, or critiques it, by selecting one of the critique options available.
4. For each critique made, the item space is narrowed down by filtering out the unwanted items, and a new recommendation is made.
5. The process continues until the customer finally selects an item.

A lot of past research has looked at critiquing in the context of high-risk, once off-items, such as digital cameras and automobiles. Since these items are highly customized and often one-off purchases, they require more effort on the part of the user to make a sound decision, since there is a larger penalty to pay if recommendation leads to a poor decision. However, research in a limited capacity has also begun to look at so-called low-involvement product domains [22]. Low-involvement product domains typically entail low-risk items, such as music and TV content. One particular work that is noteworthy in this regard is the MovieTuner feature incorporated into MovieLens, that allows movie qualities, such as *more action* to be adjusted through critiquing [23].

We propose to make use of critiquing to allow navigation of items in the *VA* space, and to gather feedback needed for computing affective offset. By allowing the user themself to take part in the recommendation process gives us feedback on how the user's perceived affective state differs from their desired state, and what they really would like to watch.

5 Recommendation Framework

5.1 General Overview

A typical system operation can be realized as follows: Once the user's mood has been detected, from audio-based parameters, the closest matching item that matches the user's mood profile is displayed to the user. The user can either accept the item, or request the recommendation of a new item. To be able to make a new recommendation, the user provides information on how the system should constrain its search. The process continues until the user finally accepts the item.

After the recommendation process has completed, the system calculates the affective offset between the initially recommended item and the finally selected item (if any), and takes this into account when processing the output labels from the classification stage, in such a way as to reflect the new mood offset. Figure C.1 shows an overview of the proposed system. We shall now present theory for the individual components.

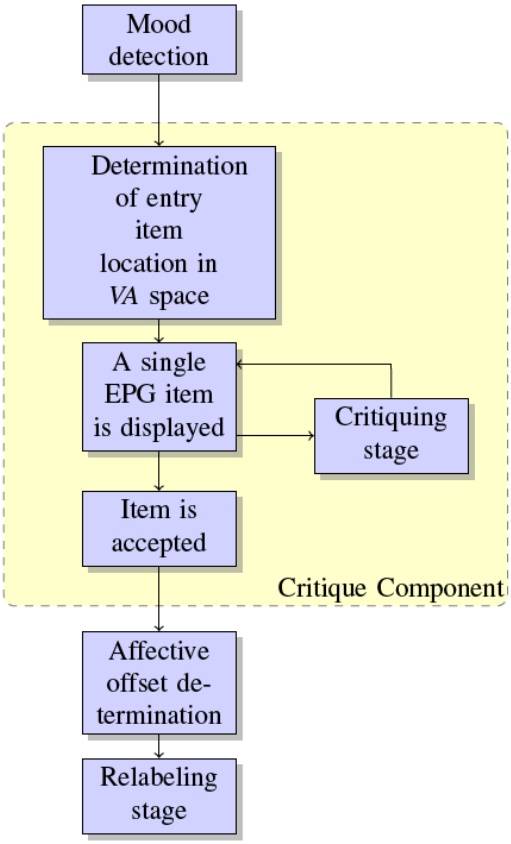


Fig. C.1: Complete system overview

5.2 Mood Detection

Since it is emotions themselves that are detectable and give rise to moods, we start by discussing emotion detection. Let E be the total number of emotion classes. Emotions can then be detected by analyzing the speech utterances from each user and assigning an emotion class $e \in E$ to each. The more classes that need to be classified, the lower the classification accuracy. What this entails is,

that for a set of utterances over a time interval for which the actual emotion was e_a, and the predicted emotion is e_p, there will almost always exist a subset of these utterances where $e_a \neq e_p$, i.e. utterances for which the actual class was not predicted correctly. What is important here is not so much that each emotions is categorized 100 % correctly, but that the areas of the emotion space, and hence adjacent emotions, that were detected, are reflected in the profile. With this in mind, the emotion profile for a single user u can be modeled by:

$$\mathbf{e_u} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_E \end{bmatrix} \tag{C.1}$$

where p_j, $0 \leq p_j \leq 1$, simply represents the actual predicted probability for emotion class j, $1 \leq j \leq E$, and $\sum_{j=1}^{E} p_j = 1$.

Over a sequence of time intervals, e.g. over the last 12 hours, the system collects the individual emotion profiles, and condenses them to a mood profile.

$$\mathbf{m_u} = \frac{1}{T} \sum_{i=1}^{T} \mathbf{e_{ui}} * w_i \tag{C.2}$$

where

$$\begin{aligned} \mathbf{e_{ui}} &= \text{The } \mathbf{e_u} \text{ corresponding to the } i^{th} \text{ time interval} \\ T &= \text{Total number of discrete time intervals, and} \\ w_i &= \text{Weighting of } \mathbf{e_u} \text{ for the } i^{th} \text{ time interval} \end{aligned}$$

To compute the weighting, a modified form of the depreciation factor, originally used in computing the depreciation citation count [24] is used to compute w_i[2]. This will ensure that emotions recorded over earlier time intervals, regardless of the size of the time interval, will always contribute less to a given overall mood profile $\mathbf{m_u}$.

The weighting w_i is thus given by the following:

$$w_i = \frac{1 + \tanh(\frac{i}{T})}{2} \tag{C.3}$$

[2]The original depreciation factor is based on years and ours is based on discrete time intervals.

5.3 Determination of Entry Item in Valence Arousal (*VA*) Space

For a given fixed set of emotions, each emotion can be characterized by associating it with an affective location (offset in degrees) around a circle[3]. There is thus a mapping from each emotion category to its corresponding angle. More formally, this set of emotions can be expressed in the following way:

$$\mathbf{\Theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_E \end{bmatrix} \tag{C.4}$$

To map the mood profile $\mathbf{m_u}$ that was introduced in the previous section, to a point in the *VA* space, we introduce the concept of a *directional mood vector*.

Each component of both $\mathbf{\Theta}$ and $\mathbf{m_u}$ is associated with a separate emotion. Therefore for each emotion j, $1 \leq j \leq E$, we create a new vector $\mathrm{Mood}_{\mathrm{VA}\mathbf{j}}$ with magnitude m_{uj} and angle θ_j, with the angle measured in degrees from the positive *valence* axis in the *VA* space:

$$\mathrm{Mood}_{\mathrm{VA}\mathbf{j}} = \begin{bmatrix} m_{uj} * \cos\theta_j & m_{uj} * \sin\theta_j \end{bmatrix} \tag{C.5}$$

This results in E separate emotion vectors, where the angle for each serves as an identification for an emotion and the magnitude indicates the confidence of that emotion, as detected by the audio classifier. Finally, all E components are summed to obtain the final directional mood vector. More formally, this is depicted as:

$$\mathrm{Mood}_{\mathrm{VA}} = \sum_{j=1}^{E} \mathrm{Mood}_{\mathrm{VA}\mathbf{j}} \tag{C.6}$$

In order to find an appropriate entry item, which forms the first stage of the recommendation process, we associate the directional mood vector with a suitable point in *VA* space. To locate the best item, we iterate through all items, where Γ is the total number of items. For each item $\gamma, \gamma \in \{1, 2, ..., \Gamma\}$ a score based on cosine similarity is computed as follows:

$$Score_\gamma = \frac{\mathbf{Mood}_{\mathrm{VA}} \cdot \mathbf{k}_\gamma}{\|\mathbf{Mood}_{\mathrm{VA}}\| \|\mathbf{k}_\gamma\|} \tag{C.7}$$

where $\mathbf{k}_\gamma$ is the location of item γ in *VA* space.

The first item to be recommended, or *entry item* is then the item γ which generates the highest score:

[3]The location of each emotion is determined by past empirical studies [9], as discussed earlier

$$\gamma = \underset{\gamma}{\operatorname{argmax}}(Score_{\gamma}), \forall\gamma, \gamma \in \{1, 2, ..., \Gamma\} \tag{C.8}$$

5.4 Critiquing Stage

At this stage the user has the opportunity to examine the entry item[4]. If he/she decides not to accept the item, a critique is specified for the new item. The possible critiques are *more pleasant*, *less pleasant*, *more intense* and *less intense*. These correspond to the affective operations more valence, less valence, more arousal and less arousal, respectively. The algorithm determines beforehand whether there is an availability of items to satisfy the potential constraint. If this condition is not satisfied, the constraint is simply not presented. Although it is possible to implement compound constraints, due to the low dimensionality of the number of free parameters available (only four), we opted for simple constraints only in this work[5].

Once the user has selected a constraint, the best matching item is determined and displayed in the following way: for a given iteration r, let S be the set of items subject to the new constraint C_r. The next item to be recommended is then the item with the shortest distance between the currently displayed item $item_c$, i.e. the last recommended item, and all other items subject to the constraint, and given as:

$$\text{Match} = \min_{\forall s \in S, s \neq c} d(\mathbf{item_c}, \mathbf{item_s}) \tag{C.9}$$

where the distance $d(\mathbf{item_c}, \mathbf{item_s})$ is a weighted form of the standard Euclidean distance in *VA* space:

$$d(\mathbf{item_c}, \mathbf{item_i}) = \sqrt{w_V * (item_{cv} - item_{iv})^2 + w_A * (item_{ca} - item_{ia})^2} \tag{C.10}$$

One of the problems with using the standard Euclidean distance is that it is based on pure distance and no consideration is given to the direction in which the user really wishes to traverse the space. Figures C.2 and C.3 show the case for a user starting out in the negative valence, positive arousal quadrant (top left), who then executes 14 critique cycles. In every case, the user selects the constraint *more pleasant*, i.e. more valence. For the unweighted case, we note that the user (unintentionally) gradually wanders over to the positive valence / negative arousal quadrant (bottom right), where, ideally, the optimum quadrant would have been the positive valence / positive arousal quadrant (top right). The weights w_V and

[4]The entry item is the very first item that is recommended to the user.

[5]Compound critiques would be suitable if the affective parameters were to be combined with other parameters, such as genre, time of day and age ratings.

w_A are therefore introduced and chosen empirically to ensure that more preference is given to either the valence or arousal dimension, depending on what constraint was chosen. This allows for a larger distance in the desired direction to be taken into consideration than would be otherwise, and results in a more direct path. The effect of using these weights is shown in Figure C.3.

The recommendation process continues until the user selects an item as acceptable, in which case it is terminated.

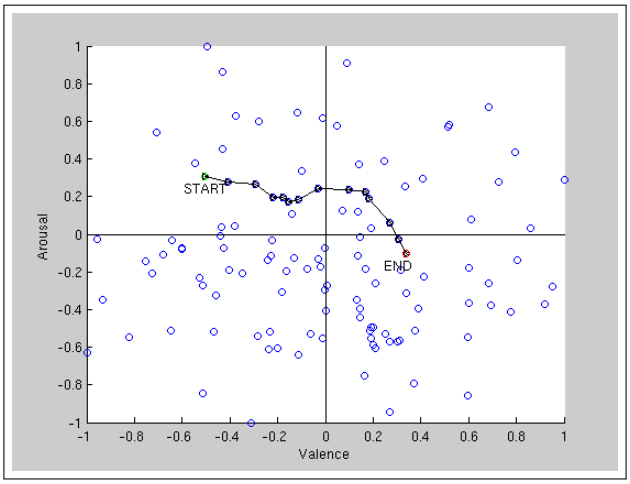


Fig. C.2: Navigating the *VA* space before the modified weighted Euclidean distance measure is introduced

5.5 Affective Offset Determination

Once the recommendation process has completed, the user will be located at another point in *VA* space. How far this point is located from the initial recommendation depends on both the number of cycles taken as well as the overall affective bearing the user took. In order to know how far off the user is from the initial recommendation, we now compute the affective offset. This offset will then be taken into account in future recommendation sessions to offset the user's mood profile (the perceived mood) with the recommended content (which relates to the desired mood).

Let A be the vector passing through the origin and the initial point where the user set out from, and let B be the vector passing through the origin and the point representing the finally selected item. The angle of this offset is given as:

$$Offset_{Angle} = \arccos\left(\frac{A \cdot B}{\|A\|\|B\|}\right) \tag{C.11}$$

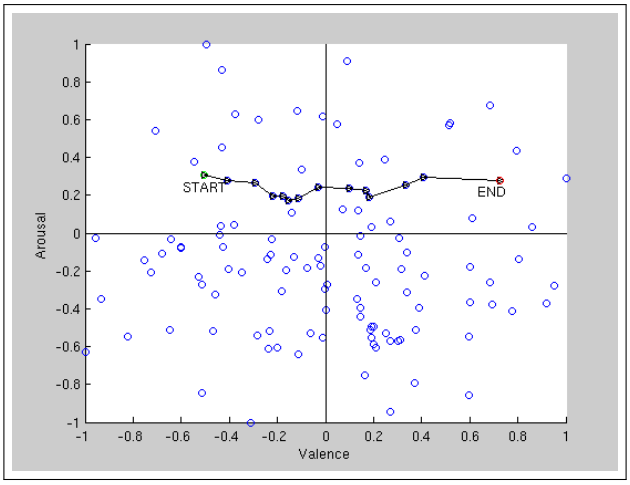


Fig. C.3: Navigating the *VA* space after the modified weighted Euclidean distance measure is introduced

where

$$\frac{A \cdot B}{\|A\|\|B\|} = \text{the cosine distance between } A \text{ and } B, \text{ and}$$
$$\arccos(x) = \theta \text{ gives } \theta \text{ in degrees and not radians.}$$

However, not only is the angle important here, but also the direction (on the emotion circumplex) of B relative to A. If we in future recommendation rounds offset the emotions in the wrong direction, instead of compensating for the mismatch between detected mood and recommended item, we would effectively be contributing to the error instead of reducing it.

We therefore determine whether this direction is clockwise, or counter-clockwise. To do this, we first compute the absolute angle of both A and B. The absolute angle for a vector through the origin (positive valence axis) to a given point $P, P = A, P = B$, is computed in the following way:

$$Angle_{P_{absolute}} = mod(-\arctan(P_y, P_x) - 90), 360) \tag{C.12}$$

where

$$\arctan(y, x) = \theta \text{ gives } \theta \text{ in degrees } (-180 \leq \theta \leq 180)$$
$$\text{mod } (\theta, 360) = \theta \text{ gives } \theta \text{ in degrees } (0 \leq \theta \leq 360)$$

Depending on the location of A and B, two possible angles can be computed:

$$Diff_c = mod(Angle_A - Angle_B, 360) \tag{C.13}$$

$$Diff_{cc} = mod(Angle_B - Angle_A, 360) \tag{C.14}$$

where

$$Diff_c = B \text{ is located clockwise relative to } A$$
$$Diff_{cc} = B \text{ is located counterclockwise relative to } A$$

If $Diff_c = Offset_{Angle}$, then this indicates that the offset occurs in the clockwise direction and $Offset_{sign} = -1$. Likewise if $Diff_{cc} = Offset_{Angle}$, then $Offset_{sign} = 1$. The sign is combined with the previously computed offset angle, to give the directional offset:

$$Offset = Offset_{Angle} * Offset_{Sign} \tag{C.15}$$

5.6 Relabeling Stage

For the 12-class emotion classifier, labels indicate the emotion that each utterance is associated with. There is no concept of distance or overlap between labels - they are simply emotion categories. However, these concepts hold for the emotional spaces themselves, and where they move, so will the labels.

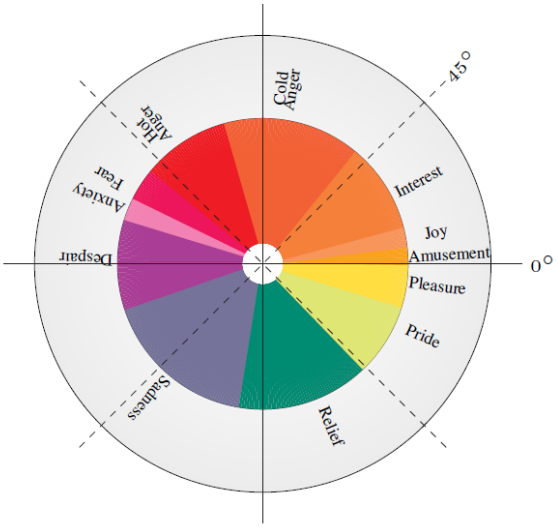


Fig. C.4: Tilting of emotion labels. By tilting the valence, arousal axis by Θ, we impose a new ordering of the labels.

Figure C.4 shows a possible configuration for a set of emotions and their location around the circumplex. For a given configuration, starting from $0°$, there exists an

	Amusement	Anxiety	Irritation	Desperation	Joy	Anger	Interest	Fear	Pleasure	Pride	Relief	Sadness
Amusement	90.00	0.00	0.00	3.33	0.00	0.00	0.00	6.67	0.00	0.00	0.00	0.00
Anxiety	13.33	23.33	3.33	6.67	3.33	6.67	3.33	10.00	6.67	10.00	0.00	13.33
Irritation	3.12	6.25	18.75	6.25	0.00	0.00	9.38	6.25	0.00	15.62	15.62	18.75
Desperation	33.33	10.00	3.33	23.33	0.00	3.33	0.00	16.67	0.00	3.33	0.00	6.67
Joy	20.00	0.00	3.33	6.67	23.33	13.33	0.00	20.00	0.00	3.33	0.00	10.00
Anger	6.67	6.67	3.33	3.33	3.33	56.67	6.67	6.67	0.00	3.33	0.00	3.33
Interest	0.00	3.33	0.00	0.00	0.00	0.00	16.67	3.33	20.00	0.00	10.00	46.67
Fear	3.33	6.67	0.00	10.00	6.67	13.33	0.00	46.67	0.00	6.67	6.67	0.00
Pleasure	3.33	3.33	3.33	3.33	0.00	0.00	13.33	0.00	43.33	0.00	6.67	23.33
Pride	16.67	0.00	6.67	6.67	0.00	13.33	0.00	0.00	3.33	40.00	3.33	10.00
Relief	0.00	10.00	0.00	6.67	0.00	3.33	10.00	0.00	23.33	3.33	43.33	0.00
Sadness	3.45	6.90	3.45	3.45	0.00	0.00	0.00	0.00	3.45	6.90	3.45	68.97

Table C.1: Confusion matrix for 12-class emotion classifier. Shaded entries correspond to actual class = predicted class.

explicit fixed ordering of the emotion labels. By tilting the valence and arousal axes by Θ, which happens to be the affective offset calculated in the previous stage, we effectively change the ordering of the labels. An important design consideration was whether to rotate the directional mood vector, as computed in equation C.6, or to rotate the labels themselves. The rationale for rotating the speech labels themselves allows for the possibility of incorporating future enrollment data, for example, as might be retrieved through multi-modal emotional systems, and leads to a better accuracy over time. Simply rotating the directional mood vector would make the system unadaptable.

Now more formally, let $L = \{l_1, l_2, ..., l_E\}$ be the set of labels. Then $X \equiv (l_1, l_2, \ldots, l_E)$ represents the sequence of labels from L before applying the affective offset. The labels in the list are arranged in order of their respective locations starting from $\Theta = \theta_1$. Likewise $Y \equiv (l_1, l_2, \ldots, l_E)$ represents the sequence of labels from L after applying affective offset, but where the list now starts from $\Theta = \theta_2$ instead. The mapping from old label l to new label is then simply carried out by the mapping function $f : L \to L$, $l \mapsto Y[Index_X(l)]$, where $Index_X(l)$ is the index of label l in X.

6 Experimental Work

6.1 Annotation of Content Items

In an initial user survey, 16 subjects rated 4 sets of 60 TV programs, with 3 subjects being assigned to each set. The TV programs were extracted from the EPG in the interval from 15:00 Friday 13 December 2013 to 10:00 on Saturday 14 December 2013. Each program shown to the evaluator was accompanied by a title, the name of the channel on which it was aired, a two-level category into which the program was placed, for example *"Level1: Movie; Level2: Comedy"*, and finally a short synopsis. All data presented to the evaluators was taken directly from the EPG metadata and was not manipulated by us in any way. The task given for each program was to read the information and thereafter rate the *pervasive mood*

of each program in the *VA* space. The method used was the well-known Self Assessment Manikin (SAM) [25], which is a psychological tool used to quantify perceived emotions. It is easy to administer, does not rely on verbs and is suitable for subjects whose native language is not English, which was the case in this study. Subjects were shown a diagram of a 9-point SAM scale, where only valence and arousal ratings were collected. It was possible for subjects to select a point anywhere on the scale, thus allowing collection of continuous valence and arousal values. Subjects were also informed that they would be rating the programs on a continuous scale. Since 3 subjects rated each TV program, a total of three ratings were collected for each. These ratings were averaged, as is customarily done [25], to give a mean SAM ratings for each program. The scales that were used can be seen in Figure C.5.

Once the rating process was complete, the first two sets were combined and the last two sets were combined, yielding two larger sets, A and B of programs containing 120 content items each. The sets were combined in this manner to create a realistically-sized number of items to browse, but taking into account the length of time required to annotate the items.

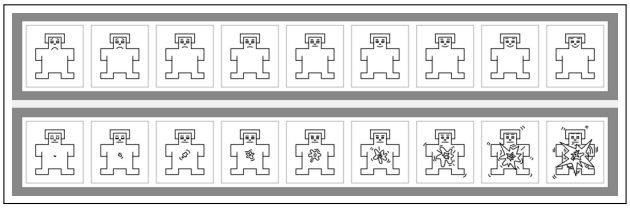


Fig. C.5: Scales used to collect the pervasive mood for each TV program. The top scale measures valence and the bottom scale measures arousal. Scales are courtesy of PXLab [26]

6.2 Mood Determination and Audio Classification of Emotions

The audio data used to represent the home user's emotional state was taken from the Geneva Multimodal Emotion Portrayals (GEMEP) [14], which was also chosen as the dataset for the emotion sub-challenge part of the Interspeech 2013 Computational Paralinguistics Challenge [27]. The dataset contains 1260 short voice utterances, divided into 18 emotional classes. The data is split across 10 actors, of which half are male and other half female. Due to the fact that 6 out of 18 of the emotions occur very sparsely in the dataset, the classification was restricted to 12 separate emotions. These were amusement, pride, joy and interest (positive valence, positive arousal), anger, fear, irritation and anxiety (negative valence, positive arousal), despair and sadness (negative valence, negative arousal) and finally pleasure and relief (positive valence, negative arousal). One of the primary reasons for selecting the GEMEP corpus was its wide spectrum of available emotions.

For each case, we connected the mood configuration to real speech utterances from the dataset by assigning each mood to the most appropriate emotions. The good mood was associated with the emotions *amusement*, *joy* and *interest*, the bad mood was associated with *cold anger (irritation)*, *hot anger*, *fear*, *despair*, *anxiety* and *sadness*, and the neutral mood was associated with the emotions *relief*, *pride* and *pleasure*. For each test trial, a speaker was randomly identified from the GEMEP dataset and a mood configuration was selected. The relevant emotion features, taken from the test set, were then concatenated and used for mood profile determination.

12-way classification of the data was carried out using front-end factor analysis (i-vectors), using the ALIZE 3.0 framework [28]. The process was as follows: 13 MFCCs (including log energy), first and second derivatives were extracted to give a fixed 39-feature frame for each 25 ms voice frame, with a 10 ms overlap for each frame. A 128-component Gaussian mixture model (GMM) was trained with the entire training set. At this point, the six unused classes were not utilized further in the system. Using the data from the GMM, a total variability matrix was trained. Subsequent to this, for each utterance, a 90-dimensional i-vector was extracted from the total variability matrix. Once in the i-vector space, classification of the utterances was then carried out using probabilistic linear discriminant analysis (PLDA) after performing normalization on the i-vectors. PLDA is known to exhibit good performance when used for the classification of i-vectors. The accuracy for the acoustic sub-system for all 12 classes on the development set was 42.72 %, and on the test set (used in the end-to-end system) was 41.20 %, which is in line with the state-of-the-art [27] [29]. More detailed results for the individual categories can be seen in Table C.1.

Emotion	Range	Mean Value
Amusement (delight)	6-12	7
Joy	5-10	7.5
Interest (activation)	20-36	28
Cold anger (irritation)	88	88
Hot anger	83-171	127
Fear	141-161	151
Anxiety (worry)	149-163	156
Despair (discouragement)	163-173	168
Sadness	144-311	227
Relief (relaxation)	249-338	293
Pride	309-3	336
Pleasure (Contentment)	347-359	353

Table C.2: Affective state locations [9]

6.3 Other System Parameters

The affective state locations used for computing the directional mood vector were adopted from past studies [9]. The range of values and calculated mean for each

	Good (M)	Good (SD)	Bad (M)	Bad (SD)	Neutral (M)	Neutral (SD)	Overall (M)	Overall (SD)
Avg iterations before (full path)	19.55	15.70	34.55	43.49	21.18	20.48	25.09	29.53
Avg iterations after (full path)	9.36	8.58	14.68	13.21	18.41	18.70	14.15	14.39
Avg iterations before (direct path)	9.64	5.49	10.36	8.44	8.82	10.29	9.61	8.21
Avg iterations after (direct path)	5.68	4.03	6.14	4.23	8.09	7.81	6.64	5.64
Ratio (direct path / full path) before	0.49	0.35	0.30	0.19	0.42	0.50	0.38	0.28
Ratio (direct path / full path) after	0.61	0.47	0.42	0.32	0.44	0.42	0.47	0.39
Affective offset (ignoring direction)	68.96	57.33	49.71	41.65	42.09	47.05	53.59	49.67

Table C.3: Summary of results for both evaluations (before and after applying affective offset) and for all three mood cases. M=Mean, SD=Standard Deviation

emotion is shown in Table C.2. For the operations *more pleasant* and *less pleasant* the weights were set to $w_v = 0.45$ and $w_a = 1$ and for the operations *more intensity* and *less intensity* the weights were set to $w_v = 1$ and $w_a = 0.45$. The user interface used for presenting the items to users and used for critiquing was implemented in PHP[6].

6.4 User Evaluation

26 subjects agreed to take part in a series of six evaluations, with each evaluation carried out on a separate day. Half of the subjects were assigned the items from group A and the other half were assigned the items from group B[7]. Each person was given access to a web portal through which they could interact with the critique-based recommender.

Each evaluation considered a single mood case. On the first day, each subject was told to picture themselves being in a neutral and relaxed mood, and to strengthen their mood, they were presented with a set of 10 neutrally rated pictures, taken from the International Affective Picture System (IAPS) [30]. The IAPS is a database of colour photographs, each annotated with valence, arousal and dominance ratings, and which is often used to elicit emotions in affective-related research studies. The subject was asked to not spend more than 15 seconds viewing each picture.

Once all pictures had been viewed, the information for a TV program was then presented to the user. The subject was asked to rate the program, on a scale of 1 to 10, on how suitable they thought the program was for the given mood.

After the program had been rated, it could either be accepted, in which case the recommendation session for that program was over, or the user could select a better recommendation case (critique the current item). If they chose to critique the item,

[6]PHP stands for "PHP: Hypertext Processor". It is a server-side scripting language suitable for the implementation of interactive web deployments.

[7]Two of the subjects had participated in the annotation phase as well - these were presented with items from the alternative group, to exclude any possibility of prior knowledge.

they were then presented with a list of choices for selecting a more pleasant, less pleasant, more intense or less intense item. For each option the number of items available for that selection were also displayed, giving the evaluator an updated indication of the potential items in each direction. In the case where no items were available, the option to navigate in that direction was not presented to the user. Furthermore, users were not prevented from navigating back over the same items they had seen before. Each new item that was presented constituted a new critiquing iteration. The number of iterations it took the user to finally select an item (by marking "accept") was counted and stored. Finally, for the final item, the user was asked once again to rate the item on a scale of 1-10, on how suitable they thought the item was.

After the subject had completed the first part of the evaluation (the neutral mood case), they were allowed to move onto the next part. Days 2 and 3 were identical to Day 1, except that different mood settings were used. For Day 2, the subject was told to instead imagine being in a good mood, where correspondingly good mood pictures were shown. In a similar fashion, Day 3 followed where the subject was now told to imagine being in a bad mood, with correspondingly bad mood pictures being shown[8]. Furthermore subjects were not allowed to complete two parts on the same day. If a subject skipped a day, a follow-up mail was sent to them. After two follow-up mails had been sent, with no response, the subject was considered to have abandoned the survey. Four of the subjects ended up dropping out of the survey, and hence we only present data for 22 out of the initial 26.

For days 4, 5 and 6, subjects were asked to repeat the evaluation for the neutral, good and bad mood cases, respectively. However, on these days, the users were not informed that their affective offset from the previous round (matching that particular mood), had been recorded and used to offset the system. Users were shown a new set of frame slides for the second round of each mood case, since seeing the same slides again would have a reduced effect.

7 Results and Discussion

7.1 Effect on the Number of Iterations

In Table C.3, we show a summary of results for both evaluations (before and after applying affective offset) and for all three mood cases. When browsing in the *VA* space to find more suitable items, users can revisit older items as many times as

[8]The following IAPS pictures were used in the evaluations: Days 1, 2 and 3, Neutral Mood: 1121, 1616, 2102, 2221, 2377, 2575, 2579, 2745.1, 7497, 7503; Good Mood: 2216, 2598, 4614, 5210, 5814, 7405, 7508, 8034, 8503, 8531; Bad Mood: 2205, 2456, 2745.2, 2751, 6313, 9185, 9252, 9635.1, 9904, 9940 - Days 4, 5 and 6, Neutral Mood: 2026, 2036, 2377, 2382, 2383, 2410, 2594, 2840, 7003, 7640; Good Mood: 1722, 1811, 2158, 7492, 8090, 8163, 8300, 8350, 8420, 8510; Bad Mood: 2399, 2682, 2683, 2703, 2800, 2900, 6021, 9420, 6570.1, 9908.

	Good (M)	Good (SD)	Bad (M)	Bad (SD)	Neutral (M)	Neutral (SD)	Overall (M)	Overall (SD)
Avg rating, Initial item, 1st evaluation	4.95	2.30	4.73	2.39	5.55	2.65	5.07	2.44
Avg rating, Final item, 1st evaluation	7.91	1.51	7.32	1.81	8.18	1.84	7.80	1.74
Avg rating, Initial item, 2nd evaluation	5.23	3.18	5.64	2.63	6.91	2.51	5.92	2.84
Avg rating, Final item, 2nd evaluation	8.91	0.92	8.32	1.17	8.73	1.12	8.65	1.09

Table C.4: Summary of ratings for both evaluations and for all three mood cases. M=Mean, SD=Standard Deviation.

they wish (in case they change their mind). Occasionally, this leads to the path from initial item to final item containing one or several loops, e.g. if while browsing, a user visits items $\{A, B, C, D, E, C\}$, the loop $\{C, D, E, C\}$ can be replaced with $\{C\}$ giving the shorter path $\{A, B, C\}$. For the sake of brevity, we refer to paths including loops as *full paths* and paths with the loops removed as *direct paths*. We are interested in these direct paths since going around in a loop essentially means the user ended up at the same spot they were at previously, and hence the same region. Direct paths are therefore a better summary of a user's ultimate migration. We therefore show results for both types of path, where the first two rows in Table C.3 show the number of iterations for full paths, and the next two rows show results for the number of iterations for direct paths. The following two rows then show the ratio between direct path and full path - the closer to 1 the ratio is, the fewer the number of loops, and the more direct the full path is. The final row shows the average affective offset for each case (ignoring the direction of the offset).

Looking at the number of iterations on average that were taken to find a suitable item, in all cases, as shown in Table C.3, we can see that a lower number of iterations was needed in the case where the user's affective offset was applied. An overall improvement was obtained of 43.60 % and for the good mood, bad mood and neutral mood cases, improvements were obtained of 52.12 %, 57.51 % and 13.08 % respectively. The average number of iterations when applying affective offset was significantly lower than when it was not applied[9] ($z = -3.39, p < 0.01, r = -0.51$). The reduction in iterations was also significant for the both the good mood case ($z = -2.90,\ p < 0.01, r = -0.44$), and the bad mood case ($z = -2.40, p < 0.05, r = -0.36$). However, for the neutral mood case, it was not significant ($z = -0.80, p = 0.42, r = -0.12$).

We believe the initial rather large standard deviation is due to the fact that browsing is rather personal. Some people generally tend to browse more than others. If browsing is indeed personal, then a Pearson correlation between the before and after iterations for each mood and all users should reveal a medium

[9]Treating the null hypothesis that the difference between the number of iterations before and after comes from a distribution of zero median, we use the sign rank test to test for significance. The effect size is computed as $r = \frac{Z}{\sqrt{N}}$ (Z is the Z-score and N is the observation count (22 users gives 44 observations)). The interpretation of r goes according to Cohen's benchmark (where a potential minus sign is ignored): $r > 0.1$ is a small effect-size, $r > 0.3$ is a medium effect size and $r > 0.5$ is a large effect size.

to large effect size. Conducting such a correlation gives values of $r = 0.28$ for the overall case, $r = 0.44$ for the good mood case, $r = 0.38$ for the bad mood case and $r = 0.18$ for the neutral mood case. The fairly strong relationship for the good and bad mood cases indicates that users are definitely more consistent in their behavior in these mood cases than in the neutral case (and more so in the good mood case).

For the direct paths, we find an overall improvement of 30.91 %, and improvements of 41.08 %, 40.73 % and 8.28 %, for the good, bad and neutral mood cases, respectively. Once again, the overall reduction was significant ($z = -2.61, p < 0.01, r = -0.39$). It was also significant for the good mood case ($z = -2.43, p < 0.05, r = -0.37$), the bad case ($z = -2.13, p < 0.5, r = -0.32$), but not significant for the neutral mood case ($z = -0.09, p = 0.93, r = -0.01$). These results indicate that even in the absence of loops, there is still a significant reduction in the path length. The higher direct path / full path ratios, for all mood cases, after applying affective offset, indicates fewer loops and more direct browsing paths.

Looking at the affective offset that arose in each case, we see the exact same trend as was seen for both full paths, direct paths, and user consistency, in terms of their statistical power. The largest affective offset was 69.96 degrees for the good mood case, followed by 49.71 degrees for the bad mood case, and finally 42.09 degrees for the neutral mood case.

These results are interesting when seen in light of the free-style user feedback comments that some of the participants provided. Four people wrote that they found it difficult to place themselves in a neutral mood setting and that the good mood setting was far easier to relate to. This might explain why in the neutral mood case there was no significant reduction in iterations - the confusing neutral mood setting resulted in participants being less consistent than in the other mood cases. The bad mood case was also considered easier to relate to, but people had more to say in general on what they thought was appropriate content for this mood. Three participants said they would only consider content that would repair their bad mood state, two wrote that comedies would be ideal, one person wrote that more intense content would be a good choice, and another two reported that if they were in a bad mood, they would not watch TV at all. It seems that the bad mood case is possibly less natural than the good mood cause and causes people to think more about what they want to watch. In the good mood case, people seem to be more open as what they want to see, and suggesting a good region allows them to more quickly find an item. In the bad mood case however, people are fussier about what they want to see - even when the region is right, more browsing is needed to find a good item.

7.2 Effect on User Ratings

In both evaluations, and for all mood cases, users were asked to rate both the initially recommended item as well as the finally selected item on a scale of 1 to 10, on how good a match they though the items were. A summary of the results for these ratings is shown in Table C.4. The first two rows cover the first evaluation before affective offset and the second two rows cover the second evaluation after affective offset.

Firstly, as expected, the final items for each evaluation were rated higher than the initial items, and in all cases these were significant: For the first evaluation, the overall increase went from 5.07 to 7.80 ($z = -5.86, p < 0.1, r = -0.88$), for the good mood case 4.95 to 7.91 ($z = -3.48, p < 0.01, r = -0.52$), for the bad mood case from 4.73 to 7.32 ($z = -3.24, p < 0.01, r = -0.49$) and for the neutral mood case 5.55 to 8.18 ($z = -3.48, p < 0.01, r = -0.53$. Likewise for the second evaluation, the overall increase went from 5.92 to 8.65 ($z = -5.82, p < 0.01, r = -0.88$), for the good mood case, from 5.23 to to 8.91 ($z = -3.44, p < 0.01, r = -0.52$), for the bad mood case from 5.64 to 8.32 ($z = -3.46, p < 0.1, r = -0.52$) and for the neutral mood case 6.91 to 8.73 ($z = -3.30, p < 0.01, r = -0.50$). This indicates that browsing was effective enough to find more suitable items.

We also looked at the ratings for only the initial item for both evaluation rounds, and found that in all mood cases, that the initial item in the second evaluation round received a higher rating that in the first evaluation round. However, in none of the mood cases was this actually significant.

More interesting though are the final ratings for both evaluation rounds. Here we found that the final ratings, after browsing had taken place, increased overall from 7.80 to 8.65 ($z = -3.54, p < 0.01, r = -0.53$), for the good mood case from 7.91 to 8.91 ($z = -2.34, p < 0.5, r = -0.35$), for the bad mood case from 7.32 to 8.82 ($z = -2.41, p < 0.5, r = -0.36$), and for the neutral mood case from 8.18 to 8.73 ($z = -1.40, p = 0.14, r = -0.21$), which were not significant. The good and bad ratings being strongly significant, and the neutral ratings not being significant suggests a link between ratings and reduction of iterations - in the neutral case users took longer to find an item they really liked (or they simply gave up), which in turn explains the low iteration reduction. The lower standard deviation for all mood cases, as noted by comparing the final ratings for both evaluations, suggests more user consensus in the higher ratings for the second evaluation than in the first. The combination of affective offset and browsing might have a stabilizing effect on users' rating behavior. We emphasize furthermore that users were not shown their previous ratings at all, and since the evaluations were carried out on separate days, would have been unlikely to recall their previous ratings. Nevertheless, in all cases we note that the final average ratings for the second evaluation were higher than *any* of the other three ratings, indicating the point of ultimate satisfaction.

The less significant initial ratings imply that applying affective offset does not necessarily help to improve the *initially* recommended item, but given the added

browsing functionality, allows a good *final* item to be located. This is an interesting finding because it indicates that single-shot recommendation of items based on users' audio features is not quite adequate. For example, a user in a good mood might be recommended an emotionally appropriate item, such as a sports game. However, if they are not interested in sports, regardless of the accuracy of the match, the item is likely to receive a low rating. It therefore makes sense to rather recommend a *region* from which the search is to be commenced, and then to harness the particular user's feedback to provide a better (more personal) recommendation the next time round.

7.3 Effect of Audio Classification

To show qualitatively how our system is affected by the inaccuracies of the audio classification component, we briefly turn our attention to six examples that show how the directional mood vector $\mathbf{Mood}_{\mathrm{VA}}$ (Equation C.6), changes with label rotations, all of which can be seen in Table C.5. To recap, a change in the configuration of emotion labels leads to a different placement of the directional mood vector, and hence determines the initially recommended item. A value of 0 indicates no rotations and corresponds to the label sequence 'amu-joi-int-irr-col-peu-inq-des-tri-sou-fie-pla'. This corresponds to observing a very low affective offset or when the finally selected item remained in the same emotion region. A value of 1 indicates one displacement and the sequence 'pla-amu-joi-int-irr-col-peu-inq-des-tri-sou-fie', a value of 2 the sequence 'fie-pla-amu-joi-int-irr-col-peu-inq-des-tri-sou' and so on. If the offset is in the other direction, the label shifting is reversed. From the figures three things are apparent:

1. Shifting of labels does not necessarily lead to an even displacement of the directional the mood vectors around the circle. This is particularly evident for the bad mood case for the female speaker.

2. Occasionally the ordering of labels is not preserved. This can be seen in the bad mood case for the female speaker and in the neutral mood case for the male speaker.

3. In some cases certain areas of the emotion space appear to be underrepresented. This can be seen in the bad mood case for the female speaker, where a large potential area for content items might be excluded.

The primary cause for these effects is due to the limited performance of the audio classifier. Since each test trial contains multiple speaker utterances, the limited accuracy of the emotion classifier causes utterances to fall into different emotion categories, which then contribute to unwanted shifting of the directional mood vector. Furthermore the emotion coordinates given in Table C.2 are not evenly spaced apart, which further contributes to the above-mentioned effects.

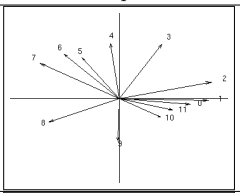
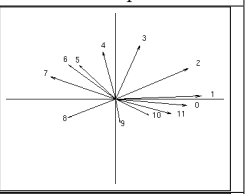
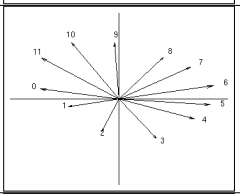
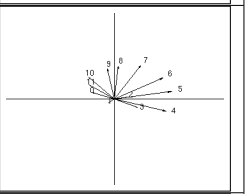
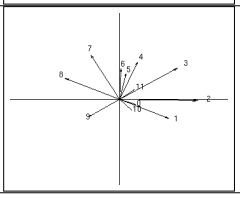
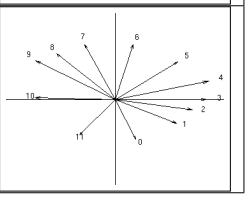
	Male Speaker	Female Speaker
Good Mood		
Bad Mood		
Neutral Mood		

Table C.5: Examples of the directional mood vector for 12 different label displacements for the three mood cases.

7.4 Limitations of the Model and Our Study

Finally, we observed five issues with the proposed model and experiments that we think are worthy of discussion:

1. The model does not handle items situated close to the VA origin very well. Take for example the case where the currently selected item is located in the positive valence, positive arousal quadrant, and where just a few browse operations leads the user to the negative valence, negative arousal quadrant. Although the distance between these items may be quite short, the resulting affective offset might be quite large. Another problem with this model is that rotation of labels will only occur when a user has moved far enough to wander into a new emotion region. For the proposed emotion offsets given in Table C.2, some areas are larger than others, meaning that more browsing will be needed to trigger a rotation.

2. As also seen in both Table C.1 and Table C.5, the effectiveness of the model is affected by the accuracy with which individual emotions can be recognized.

3. The three mood profiles for each user are assumed to be fixed. However, it is possible that some users' mood profiles might vary over time.

4. One of the problems faced with the user evaluation itself is that three of the subjects wrote that they found it difficult to browse programs in the neutral

mood setting, and that it was far easier to imagine a good or bad mood case. As evidenced by the results, this difficulty in relating to the neutral mood setting almost certainly led to the rather poor results across the board for the neutral mood setting. It appears that users perform better in a more activated mood state.

5. Two people complained that they did not necessarily always agree with the valence and intensity of programs that the initial subjects had rated, indicating just how personal each user's taste is, and also raises the question of the effectiveness of using third party annotations.

8 Conclusion

In this paper we developed a framework for recommending TV content based on moods derived from user's emotions. By allowing the user to take part in the recommendation process, we were able to compute each user's affective offset, to be used for future recommendation sessions. We used each user's affective offset to locate an initial region for recommendation, from which a recommendation was determined. The use of affective offset led to better user satisfaction overall, where ratings went from 7.80 up to 8.65. Furthermore, there was a marked decrease in the number of cycles that was needed to find a good item, compared to the case when no affective offset was applied, which went from 29.53 down to 14.39. Future work could include better modeling of items situated close to the VA origin, more predictive modeling of the directional mood vector and a framework that takes in account mood profiles that vary over time.

Acknowledgment

The authors would like to thank the Swiss Center for Affective Sciences for allowing us to use the GEMEP database. We would also like to thank Gracenotes for providing the relevant tools with which to extract the EPG data.

References

[1] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143–154, 2005.

[2] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoutic and prosodic level information fusion," *Computer Speech and Language*, vol. 27, pp. 151–167, 2012.

[3] R. Xia and Y. Liu, "Using I-vector space model for emotion recognition." in *INTERSPEECH*, 2012, pp. 2230–2233.

[4] M. Tkalcic, A. Kosir, and J. Tasic, "Affective recommender systems: the role of emotions in recommender systems," in *Proceedings of the 5th ACM conference on recommender systems*, 2011, pp. 9–13.

[5] P. Ekman, "Moods, emotions, and traits," *The nature of emotion: Fundamental questions*, pp. 56–58, 1994.

[6] P. Winoto and T. Y. Tang, "The role of user mood in movie recommendations," *Expert Systems with Applications*, vol. 37, no. 8, pp. 6086–6092, 2010.

[7] C. Peter and A. Herbon, "Emotion representation and physiology assignments in digital systems," *Interacting with Computers*, vol. 18, no. 2, pp. 139–170, 2006.

[8] J. A. Russel, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1170, 1980.

[9] N. A. Remington, L. R. Fabrigar, and P. S. Visser, "Reexamining the circumplex model of affect." *Journal of personality and social psychology*, vol. 79, no. 2, p. 286, 2000.

[10] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[11] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 566–569.

[12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech." in *Interspeech*, 2005, pp. 1517–1520.

[14] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception." *Emotion*, vol. 12, no. 5, p. 1161, 2012.

[15] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. J. Russell, and M. Wong, ""You Stupid Tin Box"-Children interacting with the AIBO robot: A cross-linguistic emotional speech corpus." in *LREC*, 2004, pp. 171–174.

[16] A. Milton and S. Tamil Selvi, "Class-specific multiple classifiers scheme to recognize emotions from speech signals," *Computer Speech & Language*, 2013.

[17] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[18] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *Proceedings of Interspeech, Firenze, Italy*, pp. 861–864, 2011.

[19] A. B. Barragáns-Martínez, E. Costa-Montenegro, J. C. Burguillo, M. Rey-López, F. A. Mikic-Fonte, and A. Peleteiro, "A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition," *Information Sciences*, vol. 180, no. 22, pp. 4290–4311, 2010.

[20] M. Z. Bjelica, "Unobtrusive relevance feedback for personalized TV program guides," *IEEE Transactions on Consumer Electronics*, vol. 57, pp. 658–663, 2011.

[21] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: An introduction*. Cambridge University Press, 2010.

[22] L. Chen and P. Pu, "Critiquing-based recommenders: survey and emerging trends," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 125–150, 2012.

[23] J. Vig, S. Sen, and J. Riedl, "Navigating the tag genome," in *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, 2011, pp. 93–102.

[24] E. Amolochitis, I. T. Christou, Z.-H. Tan, and R. Prasad, "A heuristic hierarchical scheme for academic search and retrieval," *Information Processing & Management*, vol. 49, no. 6, pp. 1326–1343, 2013.

[25] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[26] H. Irtel, "The PXLab Self Assessment Manikin Scales," 2008. [Online]. Available: http://irtel.uni-mannheim.de/pxlab/demos/index_SAM.html

[27] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 148–152.

[28] A. Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, C. Lévy, H. Li, J. Mason, J.-Y. Parfait, and U. ValidSoft Ltd, "ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2013, pp. 1–5.

[29] G. Gosztolya, R. Busa-Fekete, and L. Tóth, "Detecting autism, emotions and social signals using AdaBoost," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2013, pp. 1–5.

[30] P. J. Lang, M. M. Bradley, B. N. Cuthbert *et al.*, "International affective picture system (IAPS): Instruction manual and affective ratings," *The center for research in psychophysiology, University of Florida*, 1999.

Paper D

Audio-based Granularity-adapted Emotion Classification

Sven Ewan Shepstone, Zheng-Hua Tan, and Søren Holdt Jensen

The paper has been submitted to the
IEEE Transactions on Affective Computing.

In peer review
The layout has been revised.

Abstract

In this paper we introduce a novel framework for combining the strengths of machine-based and human-based emotion classification. The human ability to tell similar emotions apart is known as emotional granularity, which can be high or low, and can be measured. A problem with machine emotion classification, especially in the context of linking audio emotions to content items for affective search and recommendation, is that emotions may be incorrectly predicted as dissimilar emotions. For high granularity people, we propose granularity-adapted classification. Instead of identifying a single emotion class, it predicts an adapted class, allowing a larger selection of more similar items to be included. To measure the effectiveness of granularity-adaptation, we measured the emotional granularity of subjects, and for each subject, applied speech data for 12 separate emotions to two audio classifiers - one a 12-class classifier (baseline) and the other the proposed personalized granularity-adapted classifier. Using pairwise similarity judgments of emotion from each person, we could compare the most similar match for the two systems. Results show that granularity-adapted classification can improve the potential similarity by up to 28.57 %.

1 Introduction

As the quantity of available media content has exploded in recent years, a large research effort has resulted in a number of successful search and recommendation strategies that are available today and that not only locate known items, but suggest new, interesting and likable items to users. In the beginning, the focus was largely on using cognitive, data-centric approaches, such as search queries, other users' ratings to suggest movies, or meta-data based on a user's history to suggest interesting items [1]. There has also been a lot of interest in enabling machines to respond with more emotional intelligence [2]. One area that is especially interesting is that of video content retrieval. Video, being the rich content type that it is, engages the user both cognitively and emotionally, and the power of cinematographic techniques to induce emotions in viewers has long been known and exploited by directors [3]. Since emotions play such a central role in our lives, the progression to a framework for accessing video content based on emotions is no surprise, and in the field of affective computing, there have been large strides that are making emotion-related search and recommendation a reality [4].

Leveraging the power of emotions and using them in a sensible way is a challenging task and requires the seamless interplay of several key technologies. Firstly, the pervasive, or dominant emotion of the content needs to be determined. This process is usually carried out automatically using features derived from video (lighting, shots and color) or audio (energy) [5]. Secondly, the felt emotions of the user

need to be ascertained, for example using physiological measures [6] or through speech [2]. Last, but not least is the important research question of how to link users' emotions to interesting items, and how best to visualize these items in the emotional space. The IFelt system is a notable recent example of a system that allows for classification, search and retrieval of movies based on emotions, and that uses novel visualization techniques to impart emotional similarity to users [7]. Being able to accentuate to users the similarity of items from an emotional perspective, such as through a well-thought out visualization interface, can assist users to find the right items. It is precisely the interplay of these latter mentioned issues we are interested in, i.e., not only the felt emotion of the user, but how they might relate to the content to be ultimately presented. All this shall become apparent shortly.

While less accurate than other known methods, speech is probably the most practical and non-intrusive method for implicitly determining felt emotions of people. State-of-the art methods for detecting emotions from speech include traditional mechanisms such as Support Vector Machines [8], as well as modeling using i-vectors [9]. The accuracy at which individual emotions can be detected depends on several things, but is primarily determined by the system employed, the number of emotions to detect as well as whether the emotions themselves are acted out [10] [11] or spontaneous [12].

Not only does machine emotion recognition vary in accuracy, but research from the psychology domain shows that humans do not possess equal discriminating abilities when it comes to discerning emotions. The degree to which individual emotions can be distinguished is termed *emotional granularity* [13]. A person with a high emotional granularity can more easily tell emotion states apart from one another. For example, such a person would more easily discern between the emotions *irritation* and *anxiety*, which are situated fairly closely to one another in the two-dimensional valence-arousal space. A person with a low emotional granularity on the other hand tends to describe emotions in more global terms, and might tend to lump these two emotions together, as "somewhere in the negative region". High granularity users are also more consistent in their self-reports of experienced emotion over time [13].

A problem with detecting emotion classes from speech, is that certain emotions might be detected as other non-related emotions. For example, in some cases, the emotion *anger* might be mistakenly detected as the unrelated emotion *happiness*, as evidenced by the results of typical papers using audio emotion recognition [14] [15]. This is especially the case when a large number of emotion classes have to be detected. For example, results for the given dataset using two emotion classes instead of twelve, yields, as expected, a significantly higher unweighted average recall [11]. Similarity from an audio feature perspective does not necessarily imply similarity in the emotion domain. As far as recommendation is concerned, this could be a problem if content items to be recommended to a user fall within a region in the valence-arousal space that has been mapped to the incorrectly-predicted emotion. Even if the region is correctly selected, a further problem is that the region's limited

size limits the number of potential items that can be recommended. Due to the fuzzy nature of both emotion classes and content items rated from an emotion perspective, having a larger region would increase the likelihood of including more similar, and likable items to the user.

One possible workaround to these problems is to lower the number of emotion classes to be detected. This can be done by defining a smaller number of *adapted classes*, where an adapted class is no longer limited to a single emotion category, but may contain multiple, similar emotion categories. Having a lower number of such adapted classes allows more training data per adapted class, is an easier task for an audio classifier to handle and reduces the risk of incorrect predictions. Once an appropriate adapted class has been identified, it can be mapped to a larger region, from which more interesting content items can be drawn. For high granularity people, we postulate that this would be beneficial to them, since as they can more easily distinguish the subtle emotions in the region from one another, selection of the final item (the most liked item given their current emotion) can be left up to them. It is in this way that we propose to combine human and machine emotion recognition - the users that are more adept at distinguishing emotions can be assigned larger selections of items.

The main contribution of this paper is a novel method that utilizes the idiosyncratic emotional granularity of tested subjects to ultimately allow for a better match between emotions extracted from speech and emotion-labeled items, such as movies. Since high granularity users are assumed to have less difficultly in telling similar emotion states apart anyway, it is possible these users do not need high-resolution emotion classification. We propose therefore to utilize the knowledge of granularity to alter an emotion classifier in an attempt to bring together limits in machine detection with strengths in human detection. We call this *granularity-adapted emotion classification*, and propose it as an audio front-end to more advanced search or recommender systems. The main hypothesis is that granularity-based emotion classification can be used to predict adapted classes that include more similar emotions, when compared to a test utterance, than single classes. An additional contribution of this paper is a state-of-the art emotion classifier that is built from the fusion of an i-vector system and an SVM system.

The remainder of this paper is structured as follows: Section 2 discusses emotion theory, particularly that of emotion granularity in people. Section 3 then presents an overview of automatic detection of emotions in speech. The following section introduces the framework for determining the suggested granularity-adapted classes, and is the main contribution of this paper. Section 5 discusses experimental work and here we also present a state-of-the-art fusion system incorporating two sub-systems. The following section presents our results and findings, and the final section concludes the paper.

2 Emotional Granularity

Probably the most well-known way for modeling emotions is to represent discrete emotional states as regions in the two-dimension space spanned by valence and arousal, which we will henceforth refer to as the *VA* space. Valence refers to the hedonic quality or pleasantness of the emotion experienced, whereas arousal refers to the perception of arousal (also called activation or intensity) of the emotion. Russel's Circumplex Model of Affect [16] [17], extensively studied in psychology research, depicts emotional states as occupying the periphery of a circular structure.

The amount of emphasis a user places on either valence or arousal can be portrayed in geometric space [13]. In the ideal, prototypical case, a user will weight the valence and arousal dimensions equally, and the circumplex will form a circular structure. This would be the case for a high granularity person. However, as is more commonly the case, a user will place more weight on either the valence or arousal dimension (typically valence more than arousal), resulting in a more squashed or elliptical circumplex structure. Valence focus refers to the extent to which a user is able to tell emotions apart on the valence scale. Arousal focus instead refers to the extent to which a user weights the arousal dimension.

In this work, we define valence focus as ϕ_{vf} and arousal focus as ϕ_{af}, both of which can be computed for any individual. In Section 5 we give an overview of how ϕ_{vf} and ϕ_{af} can be computed.

There have been other approaches in the literature to express emotion granularity. In terms of the taxonomy of emotions, granularity can be expressed in terms of the hierarchy of emotions, and can be expressed at different levels [2]. At the highest level the categorization may simply be positive and negative. For example, negative and non-negative emotions have been detected before from speech data obtained from a call center application [18]. At the next level follows individual emotion categories, such as *pleasure*, *anger*, and *fear*. At the lowest level are sub-categories, such as *hot anger* and *cold anger*.

One particular study discusses granularity, but in the context of annotation, where multiple emotion granularities can be represented [19]. The authors present a framework for hierarchical annotation called Multi-level Emotion and Context Annotation Scheme (MECAS), where two emotion labels per segment of speech can be specified: a major, dominant emotion and a minor, background emotion.

Yet another work on granularity looks at improving recognition of prosodic events by augmenting audio features with Parts of Speech (POS) feature flags [20]. 15 POS features for each current, previous and following word and define the smallest context size. To reduce computation complexity in training the Multi-layer Peceptron (MLP), a different granularity is employed where the 15 POS classes are reduced to 6 cover classes.

3 Automatic Emotion Recognition from Speech

Emotion classification in speech is a challenging task and has received a lot of attention in the past ten years. While there is recent interest in continual modeling of emotions [21], speech utterances are generally assigned to fixed labels, such as Ekman's "big six" emotions (*anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*) [22], and emotion speech datasets (corpora) typically contain either acted speech [23] [10] or spontaneous speech [12] assigned to fixed emotion labels. The two major feature types are acoustic and linguistic features. Linguistic features are of more value when the speech is spontaneous and not based on any pre-defined script, and acoustic features are more applicable for acted databases [2]. When acoustic features are used, automatic emotion classifiers have been observed to perform better with respect to classifying arousal than valence [11]. Although the majority of earlier papers focused on emotion recognition derived from acoustic information, combining this acoustic information with other sources of information, primarily at the language level, has shown improvement in performance [18].

After any necessary speech-signal pre-processing, low-level feature descriptors are extracted, from which an appropriate model can be constructed. Many parameters are used to detect emotion, including mel-frequency cepstral coefficients (MFCCs), which have been the most investigated features for emotion recognition. In addition to spectral features such as MFCCs and formats, prosodic features such as pitch, intensity, duration and to a less extent, voice quality features, are also used [2]. Furthermore, to take advantage of some of the longer term phenomena of emotions that can occur over many frames, the modeling can be enhanced by mapping of variable length features to fixed length vectors, using functionals (so-called static feature modeling).

Emotions are modeled using a wide variety of techniques including generative models such as Gaussian mixture models (GMMs), as well as discriminative models such as support vector machines [8], which provide good generalization performance, and back propagation artificial neural networks (ANNs). A recent method for modeling emotions include anchor models based on Euclidean and cosine distance metrics, which are used as a feature extractor to enhance emotion recognition [24]. Here, test utterances are compared to emotion classes in the anchor space. Another method models emotions using front-end factor analysis (i-vectors) [9], [25], which are currently considered state-of-the-art in speaker recognition.

4 A framework for Determining Granularity-adapted Classes

4.1 Localization of Emotions in *VA* Space

In this work, we shall represent emotions as coordinates in the two dimensional valence and arousal Euclidean space. As seen in the previous section, Euclidean metrics to measure the similarity of emotions has been applied before, albeit in another context [24]. Let E be the maximum number of emotion classes available. Based on a given person's similarity ratings for all pairs of the E emotion terms, it is possible by using non-metric multidimensional scaling (MDS) to render these terms as coordinates in an N-dimensional geometric (Euclidean) space. When mapped to a two-dimensional space, the Euclidean distance between any two emotion terms is inversely proportional to the similarity rating for that same pair. This means that for two emotions that are very similar to one another, they will be situated close to each other. Furthermore, a visual inspection of the new coordinates in this 2-dimensional space meaningfully reveals the underlying dimensions as being valence and arousal. It is also possible to compute emotion coordinates for an entire group of users using individual differences multidimensional scaling. We define the list of E emotion coordinates for such a group by the sequence $M = \{\mu_1, \mu_2, ..., \mu_E\}$.

4.2 Determining each User's Valence and Arousal Focus

For each user we can determine a valence focus ϕ_{vf} and arousal focus ϕ_{af}. For a given similarity rating of two emotions, we wish to determine how much these ratings account for the correlation between ratings of the experience of these same two emotions [13]. Each focus value can be computed by correlating a distance matrix, obtained for a group of user's pairwise similarity ratings, to correlations obtained from individual self-report ratings of experienced emotion. For example, if two emotion terms are similar in valence, but different in arousal, and this fact is not seen in correlations obtained from self-report ratings for these two terms, this will result in a low valence focus. A more detailed account of how these were computed in this work is given in Section 5.

4.3 Granularity-based Class Adaptation

When performing audio classification, for a given user u and a sequence of speech utterances, the user's emotion profile can be expressed as:

$$\mathbf{e_u} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_E \end{bmatrix} \tag{D.1}$$

where p_j, $0 \leq p_j \leq 1$, represents the actual predicted probability for emotion class j, $1 \leq j \leq E$, and $\sum_{j=1}^{E} p_j = 1$.

The emotion class $e \in E$ that is assigned is usually that class with the highest likelihood. The more classes that need to be classified the lower we expect the accuracy to be. Let us assume that each class corresponds to a single emotion category (this is usually the case in emotion classification). For low granularity subjects, who are presumed to have difficulty in telling similar emotion states apart, the best we can do is to predict the correct emotion from the audio utterance, and assign one or more content items from a potentially small region in the valence arousal space that is indicative of that emotion. For example, if the emotion has been detected to be *pleasure*, then the region that more or less corresponds to that particular class would be selected, and all content items found to match this region could then be included in the list of items to be recommended.

High granularity subjects, however, can more subtly tell emotions apart and we propose to use this fact to our advantage. Here the proposal is that instead of trying to classify a large number of discrete emotion classes, each tied to a single emotion, we focus on a smaller number of adapted classes, where the assumption is that each adapted class is allowed to contain multiple (similar) emotions instead of just one emotion. Having a lower number of such adapted classes to select from ought to increase the accuracy with which each can be detected. By once again mapping each adapted class to a region in the valence arousal space, the adapted classes containing more emotions will ultimately lead to larger regions, meaning that a larger number of content items (more than for the discrete emotion case) can be recommended.

By presenting high granularity users with this larger list of content items, that all fall within the selected emotion region, allows for the selection of more suitable items. In this work, we are not overly concerned with exactly what the definition of *suitable* should be; instead here we focus on improving the audio classifier that would allow for the extraction of a larger number of more similar items for the subject's perusal. Whether the items that are presented exactly match the emotion detected or are tailored to what individual subjects prefer for a given emotion (e.g. showing good mood items for people in a bad mood), should be treated as contextual information, and be left up to the recommender's content filtering algorithm to decide [4]. From a presentation and visualization standpoint, items belonging to the same adapted class could be presented in a two-dimensional plane, possibly using different colors to mark individual adapted classes, as has been attempted before [7]. This manner of displaying items, as opposed to using a simpler structure such as a list, would also allow users to more easily discover unknown items, since items that are emotionally similar to one another would lie in close proximity to one another.

Using the distance between the set of E emotions, we can capture the particular arrangement in a dendrogram. A dendrogram is a well-known structure used to depict the arrangement of clusters and eloquently captures the hierarchy of the clustering. In this work we shall simply equate a cluster to a set of utterances having the same emotion label (class). In Figures D.1 and D.2 below, we show a

set of 12 emotion classes and the corresponding dendrogram, based on the MDS Euclidean distance between them.

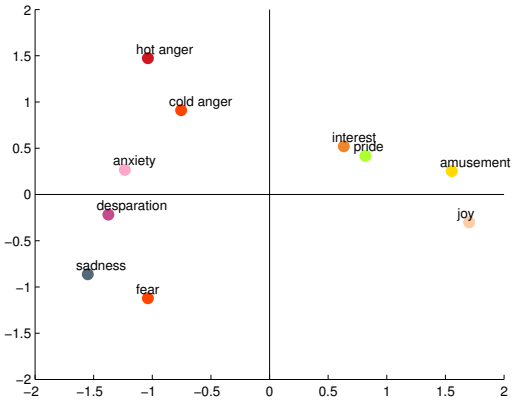


Fig. D.1: An example of individual stimulus coordinates extracted from an MDS analysis. The arrangement of the emotions meaningfully reveals the dimension along the X-axis to be valence and that along the Y-axis to be arousal.

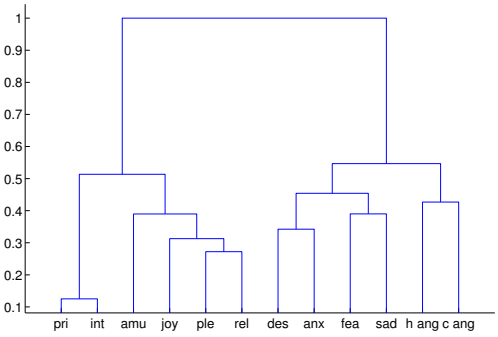


Fig. D.2: A MATLAB-generated dendrogram constructed from the emotion coordinates. Heights of linkages normalized to 1.

In a dendrogram, nodes are connected together as linkages, and each link results in two clusters being joined together. The height of each node (emotion class) is proportional to the distance, or dissimilarity between its children, and this is what gives the dendrogram its hierarchical levels. For low heights, there are a large number of clusters that have not yet been merged and as the height increases, more and more clusters are merged, starting with more similar clusters being merged first. Therefore each height value is a cut-off point that corresponds

to a unique clustering arrangement. Our proposal is to use individuals' granularity as a key to indexing the height in a dendrogram, and in so doing, obtaining a clustering arrangement that is proportional to granularity: The larger the overall granularity of the user, the more emotion classes per cluster and the fewer the number of clusters overall. In other words, applying the granularity (valence and arousal focus) information can be seen as a pruning operation of a tree structure - all objects below each cut constitute a single cluster. Shortly, we shall use this information to derive the number and contents of the granularity adapted emotion classes.

Since each individual's granularity is characterized by both valence focus and arousal focus, we construct a separate dendrogram for each, where instead of using Euclidean distance as the dissimilarity metric, we use the respective valence and arousal one-dimensional distance values. This results in a valence dendrogram, derived from the valence-based clustering arrangement, and an arousal dendrogram, derived from the arousal-based clustering arrangement.

Given a set of one-dimensional stimulus coordinates (valence or arousal only), Algorithm D.1 is used to return the dendrogram D that links the height values to a given clustering arrangement:

Algorithm D.1: ConstructLinkages

input : *stimCoords*
output: Dendrogram structure D
`// Get pairwise distances between pairs of the stimulus coordinates`
$P = pdist(stimCoords)$
`// Compute linkage heights`
$D = linkage(P)$

We normalize the height values, corresponding to the points at where clusters are linked to one another, to a value between 0 and 1:

$$linkages_{Norm} = \frac{linkages}{\max(linkages)} \tag{D.2}$$

Likewise, focusing just on valence for the time being, for each user's valence focus, we normalize it to a value between 0 and 1:

$$\phi_{vf_{Norm}} = \frac{\phi_{\mathrm{vf}} - \min(\phi_{\mathrm{vf}})}{\max(\phi_{\mathrm{vf}})} \tag{D.3}$$

The index into the height of the dendrogram can then be computed using Algorithm D.2. Using this height as the cut-off value, we obtain a valence-based clustering, which gives a suggestion for the formation of the adapted classes, but with reference to valence only.

Algorithm D.2: ComputeAdaptedClass

input :
output: Adapted Class C
D = dendrogram of linkages, from Algorithm D.1
$vfNorm$ = normalized valence focus, from (D.3)
$LNorm$ = list of normalized linkage distances, from (D.2)
for $i = 1$ **to** $LNorm$ **do**
  if $vfNorm >= i$ **then**
    $vfHeight = i$ break
  end
end
`// Get all objects below cut off height`
$C = cluster(D, vfHeight)$

We now proceed to derive the adapted classes. Let N_v represent the number of unique valence clusters, as read off from the valence dendrogram, let the sequence $CV = (cv_1, cv_2, ..., cv_{N_v})$, represent the sorted (at the cluster level) list of these valence clusters (their ids), and let $V = (V_1, V_2, ..., V_{N_v})$ be the same list, but with elements in each cluster replaced by their valence values (as defined in *VA* space). By sorting, we imply that for any two elements $cv_{k,x}$ and $cv_{k+1,y}$ between two adjacent clusters k and $k+1$, that $\max(V_k) < \min(V_{k+1})$. This gives an ordering at the cluster level, while disregarding the order of elements within a particular cluster, and can be seen as a bucket sort, without sorting within the buckets themselves. Note that each cluster cv_k will contain one or more elements (emotion classes), and that an element cannot reappear throughout the sequence CV. Furthermore, there is no constraint imposed that clusters have to hold the same number of elements (although all individual emotion classes must be accounted for across all clusters).

This process is repeated using the arousal focus of each individual and arousal dendrogram, to determine N_a, the number of unique arousal clusters, CA and A, the sorted list of arousal cluster ids and their arousal values, respectively.

Armed with these valence and arousal clusters, of which some will contain multiple elements, we partition the valence axis into N_v bands. (Apart from the lower bound for the first band and the upper bound from the last band, which are -1 and 1 respectively, the imaginary boundary for two adjacent bands, k and $k+1$ could be defined as $\frac{\min(V_{k+1}) - \max(V_k)}{2}$). Likewise we divide the arousal axis into N_a bands. For elements that are common to the area where two bands cross, we combine them to form the granularity adapted classes. The theoretical maximum number of possible granularity adapted classes is given by $T = N_v * N_a$. However, the actual number of adapted classes will in fact usually be much lower, since many of the regions where the valence and arousal bands cross will not contain

any elements, and will thus be empty. For each region, where two bands cross, the corresponding adapted class is denoted by $\Psi_{i,j}$, which contains elements from valence band i and arousal band j and is computed by:

$$\Psi_{i,j} = CV_{\mathrm{i}} \cap CA_j, \forall i, i \in \{1, 2, ..., N_v\}, \\ \forall j, j \in \{1, 2, ..., N_a\} \tag{D.4}$$

4.4 Assigning of Enrollment Labels

Each adapted class $\Psi_{i,j}$ corresponds to a separate region in *VA* space. Within this adapted class, there is a one-to-one mapping between each element (indivisible class) and its audio classification label. The mapping in this set from each element p, $\forall p$, to its physical label e_k is carried out by the mapping function $f : M \to E$, $\mu_k \mapsto E[Index(p)]$, where $Index(p)$ returns the index k of the element μ_k from M that matches p.

4.5 Linking Adapted Classes to Regions in *VA* Space

The location of content items in the valence arousal space, such as movie items, is customarily obtained by means of manual annotation by multiple evaluators [26] [27], often using a non-verbal rating method such as the Self Assessment Manikin (SAM) [28]. This process can also be carried out automatically, as outlined in the beginning of this paper [5].

Once a granularity adapted class containing one or more adjacent emotion categories has been identified, it needs to somehow be related to the continuously labeled content items. One possible way to achieve this is by fuzzy-clustering of the two-dimensional content items into discrete clusters, and then using a membership function to relate each content item to one or more physical emotion categories. Such a scheme has been proposed before in the context of automatic video indexing [29]. Another method is to compute a score based on the cosine similarity between a vector representing the mean location of the adapted class and content items [30]. The resulting emotion categories are then just the adapted classes (sets of labels). Furthermore, soft counts could be used to determine the most applicable set of content items for a given adapted class (for example, all items whose proportional membership to a set of classes falls within a certain window). If on the other hand, movies have been tagged directly using discrete emotion categories, the process of linking adapted classes to sets of content items is more straightforward. Finally, the user who is presented with content items from a region matching the proposed adapted class can then make the final selection.

5 Experimental Work

	Amusement	Joy	Pride	Desperation	Hot Anger	Fear	Anxiety	Cold Anger	Sadness	Interest	Pleasure	Relief
Amusement	50.00	6.67	3.33	10.00	3.33	0.00	0.00	6.67	6.67	10.00	3.33	0.00
Joy	10.00	30.00	0.00	6.67	0.00	3.33	0.00	6.67	0.00	23.33	3.33	16.67
Pride	0.00	3.33	30.00	16.67	0.00	6.67	6.67	10.00	0.00	6.67	10.00	10.00
Desperation	6.67	3.33	0.00	56.67	3.33	0.00	0.00	0.00	13.33	6.67	10.00	0.00
Hot Anger	6.67	3.33	3.33	0.00	36.67	6.67	13.33	20.00	10.00	0.00	0.00	0.00
Fear	3.33	3.33	3.33	0.00	20.00	40.00	3.33	3.33	13.33	0.00	10.00	0.00
Anxiety	6.25	3.12	9.38	0.00	3.12	3.12	28.12	12.50	21.88	12.50	0.00	0.00
Cold Anger	10.34	3.45	3.45	0.00	6.90	0.00	0.00	68.97	3.45	3.45	0.00	0.00
Sadness	20.00	3.33	0.00	0.00	10.00	3.33	3.33	20.00	36.67	0.00	3.33	0.00
Interest	6.67	10.00	0.00	10.00	0.00	0.00	3.33	0.00	3.33	56.67	6.67	3.33
Pleasure	0.00	20.00	3.33	16.67	3.33	0.00	6.67	6.67	0.00	20.00	6.67	16.67
Relief	10.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.67	3.33	76.67

Table D.1: Confusion matrix for 12-class emotion classifier. Shaded entries represent the results where actual class = predicted class.

5.1 Audio Classification of Emotions

The audio data used to represent the user's emotional state was taken from the Geneva Multimodal Emotion Portrayals (GEMEP) [10], which was also chosen as the dataset for the emotion sub-challenge part of the Interspeech 2013 Computational Paralinguistics Challenge [11]. The dataset contains 1260 short voice utterances, and we used the same partitioning in our system as was specified for participants for the challenge. Out of the total of 1260 utterances in the entire set, 482 were allocated for training, 236 were allocated for system validation, and 362 were used for testing purposes. The data is divided into 18 emotional classes, and split across 10 actors, of which half are male and other half female. Due to the acted nature of the database, we consider acoustic features only. Since 6 out of the 18 emotions occur very sparsely in the dataset, the classification was restricted to 12 separate emotions. These were *amusement*, *pride*, *joy* and *interest* (positive valence, positive arousal), *anger*, *fear*, *irritation* and *anxiety* (negative valence, positive arousal), *despair* and *sadness* (negative valence, negative arousal) and *pleasure* and *relief* (positive valence, negative arousal). One of the primary reasons for selecting the GEMEP corpus was its wide spectrum of available emotions.

In order to classify the emotion data an i-vector-based system (state-of-the-art in speaker recognition) as well as a standard SVM classifier (a popular choice for emotion recognition) were used. Due to the large number of emotions that need to be classified and the low degree of emotional prototypicality of the dataset, the overall accuracy for each system is fairly low, and therefore classification was carried out by fusing the results from each system.

In both systems the number of classification classes was set to 12 for the baseline case, and in the adapted case, uniquely determined from each user's valence and arousal focus. The adapted classes and their labels for the adapted classifier for each individual was determined as follows: Using the group-wise stimulus coordinates (a single set), and each individual's valence focus, we used Algorithm D.1, (D.2), (D.3) and Algorithm D.2 to determine the relevant cut-off value and from

that obtained an appropriate valence clustering[1]. The same was done to determine an arousal clustering. Finally, the valence and arousal clusters were combined to form granularity adapted classes according to (D.4). In the adapted case, all emotion labels corresponding to the same granularity adapted class were given the same supervised label before retraining the classifier.

The i-vector system was constructed as follows: 13 Mel Frequency Cepstral Coefficients (MFCCs) (including log energy), first and second derivatives were extracted to give a fixed 39-feature frame for each 25 ms voice frame, with a 10 ms overlap for each frame. MFCCs are simply a compact representation of the spectral envelope of a speech signal. A 256-component Gaussian mixture model (GMM) was trained with the entire training set. After this, the six unused classes were not utilized further. Using the data from the GMM, a total variability matrix was trained. After this, for each utterance, a 150-dimensional i-vector was extracted from the total variability matrix. Once in the i-vector space, classification of the utterances was then carried out using probabilistic linear discriminant analysis (PLDA) after performing normalization on the i-vectors. The accuracy for the i-vector sub-system for all 12 classes on the test set was 40.31 %.

The Multi-class SVM system was constructed using the 2013 ComParE feature set, which contains 6373 features. The configuration for extracting these features is publicly available in the OpenSmile toolkit [31]. The LibSVM toolkit [32] was used to train the multi-class classifier. We trained the SVM using probability estimates (to allow for easy fusion with the i-vector system), a cost C of 0.07 and a gamma value g of 0.1. The accuracy for the SVM system was 38.40 %.

Scores from both systems were combined using weighted summation based fusion, which resulted in a higher overall system accuracy for the baseline case of 43.03 %, which is in line with the state-of-the-art [11] [33]. Table D.1 shows the 12-class baseline results for the combined system.

5.2 Computing of Individual Valence and Arousal Focus

Computing a subject's granularity required them to take part in two surveys. The first evaluation involved determining the semantic similarity of 12 emotion categories, taken from the GEMEP database, that are more or less evenly spaced around the affective circumplex. This resulted in $\frac{12!}{(12-2)!(2)!} = 66$ possible pairs (without repetition) that had to be rated. For each pair, presented in random order, subjects were asked to give a rating of 1 to 7 on how similar the emotions in the pair were, with 1 being extremely dissimilar, 4 unrelated, and 7 extremely similar.

[1]This was carried out in MATLAB using the Statistics Toolbox. The functions *pdist*, *linkage* and *cluster* are all standard commands in this toolbox.

In the second evaluation, subjects were shown 24 slides[2] from the International Standard Picture Database (IAPS) [34]. The IAPS is a database of colour photographs, each annotated with valence, arousal and dominance ratings, and which is often used to elicit emotions in affective-related research studies. For each slide, they were presented with all 12 emotion categories (in random order) and asked to rate the slide for *each* category. The question posed was: "To what extent does this slide invoke <*emotion*> in you?". The rating scale given to subjects was 1 (not at all), 3 (moderate amount), and 7 (a great deal).

To determine each subject's valence focus and arousal focus, we adopted methods used from previous studies in psychology [35] [13]. First, the semantic similarity ratings obtained from the first evaluation were subjected to a weighted individual differences scaling MDS procedure (INDSCAL) in SSPS [36][3], from which a set of 12 emotion *stimulus coordinates* was obtained in two dimensional space[4]. The distance along the valence dimension between each emotion pair (of which there were 66) was entered into a single valence distance matrix. The same was done in the arousal dimension to create an arousal distance matrix.

From the picture data a 12 by 12 correlation matrix was constructed for each person, once again representing 66 correlations, e.g *sadness-pride*. This matrix captures the relatedness in each individuals emotional reports of experience of the IAPS slides. After carrying out Fisher's *R-Z* transformation on the data, the individual pairs were correlated with the valence distance matrix to yield an index for valence focus ϕ_{vf}. By carrying out such a correlation, it is possible to determine just how much of the granularity in the experience of two emotions is due to emphasis on valence and arousal. Similarly the correlation matrix for each person was correlated against the arousal distance matrix to give the arousal focus ϕ_{af}. Since the distance matrix is inversely proportional to the correlations from the picture data, the values are inverted.

5.3 Evaluation of Granularity Adapted Classification

Using each subject's valence and arousal focus, we derived the adapted classes needed to implement the granularity adapted classifier. The first two columns of Table D.2 shows each user's valence and arousal focus, and the following three columns give values for the number of valence and arousal clusters, and combined

[2]The following IAPS pictures were used: 1710 Puppies, 2050 Baby, 3110 Burn Victim, 3185 Stitches, 4700 Couple, 5000 Flower, 5301 Galaxy, 5621 Sky divers, 5760 Outdoors, 5920 Volcano, 6230 Aimed gun, 6415 Dead Tiger, 6825 Military, 6832 Police, 7001 Button, 7080 Fork, 7010 Basket, 7023 Garbage, 7234 Ironing board, 8160 Man on cliff, 8190 Skiers, 8461 Happy teens, 8470 Gymnast, 9001 Cemetery.

[3]SSPS requires the data to be entered as *dissimilar* ratings.

[4]SSPS produces a stress vs dimension plot, where an elbow on the plot at a stress value of 0.18 indicates that two dimensions are the best fit. A high squared correlation $R^2 = 0.87$ indicates that a large proportion of the variance in the MDS solution is accounted for by the similarities between the rated emotion words.

Subject Id	Valence Focus ϕ_{vf}	Arousal Focus ϕ_{af}	Valence Clusters	Arousal Clusters	Adapted Classes	Accuracy adapted %	Similarity Baseline	Upper bound similarity: Adapted	Lower bound similarity: Adapted	Mean similarity: Adapted
1	1.02	0.07	11	4	11	43.51	1.04	1.02	1.09	1.05
2	1.78	-0.07	1	11	11	42.00	1.06	1.08	1.20	1.14
3	1.31	-0.05	1	11	11	43.94	1.03	1.04	1.15	1.10
4	1.79	-0.01	1	7	7	43.57	1.06	0.87	1.77	1.34
5	1.49	0.03	1	6	6	42.04	1.03	0.85	1.99	1.49
6	1.40	0.01	1	6	6	42.04	1.08	0.80	1.99	1.45
7	1.89	0.05	1	5	5	50.43	1.02	0.77	1.68	1.24
8	1.56	0.05	1	5	5	50.43	1.01	0.74	1.80	1.31
9	1.57	0.06	1	4	4	53.01	1.00	0.46	2.57	1.56
10	1.41	0.06	1	4	4	42.00	1.05	0.66	2.43	1.54
11	1.12	0.21	4	1	4	45.66	1.08	0.84	1.97	1.39
12	1.37	0.09	1	3	3	58.50	1.06	0.46	2.63	1.60
13	1.41	0.10	1	3	3	58.50	1.11	0.27	2.70	1.63
14	1.14	0.28	3	1	3	51.42	1.05	0.70	1.94	1.36
15	1.18	0.18	2	1	2	63.69	1.07	0.61	2.46	1.58
Mean	1.43	0.07	-	-	-	53.87	1.05	0.75	1.96	1.38
SD	0.26	0.09	-	-	-	14.41	0.03	0.23	0.53	0.19

Table D.2: Main results: Valence Focus(Z-score), arousal Focus(Z-score), number of valence clusters, number of arousal clusters, number of granularity adapted classes, adapted classifier accuracy, best attainable similarity, worst attainable similarity, mean similarity. Shown for all users including the mean and standard deviation (where applicable).

granularity regions.

In terms of the feasibility of using such an audio classifier in the proposed framework, there is the advantage that an individuals valence and arousal focus need only be computed once. This implies that determining the adapted-classes for a given individual and training the classifiers is a once-off process. Furthermore, for individuals in emotional granularity that lies below a predefined threshold, it might be decided to just utilize the baseline classifier instead.

For each test trial from the GEMEP dataset, two classifications for each user were carried out: The first classification was a standard 12-way classification of the selected emotion data, and was implemented using the baseline classifier described in the previous section. The division of classes was therefore not linked to the subject's granularity in any way. Thus for each emotion utterance in the test set, a resulting single emotion class was predicted. This resulted in 362 <*actual class, predicted class*> pairs. The second classification was the adapted classification. For each utterance in the test set, a single adapted class, containing 1 or more emotions, was predicted, which resulted in another 362 <*actual class, predicted adapted class*> pairs. The same test data was reused for each of the 15 subjects.

To analyze the effectiveness of granularity-based adaptation, we used the already-provided similarity data from each user to compare each user's perceived similarity of emotions between the baseline and adapted classifier. Similar to the group MDS scaling procedure, we extracted similarity stimulus coordinates for each individual based on their ratings, and used these as the ground truth for evaluating the similarity of emotions for each individual. From this data, it is possible to tell just how far apart two emotions are by examining their distance in the Euclidean space. We define the best attainable similarity as an upper bound for the most similar

items that can be included. By computing the best attainable similarity between the actual and predicted classes for each test utterance, we develop a notion for how similar, from a similarity of emotions perspective, subjects might regard the classification.

In the baseline case, we compare the similarity between the labeled emotion of the incoming utterance and the resulting single prediction (i.e., a single emotion class). In the adapted classifier case, where subjects have the ability to pick out the most applicable item in a region, we compare the similarity between the incoming emotion label (utterance) and both the adapted class's *most similar* and *least similar* emotion. The most similar emotion gives an upper bound on the maximum theoretical similarity that can be attained and the least similar gives the corresponding lower bound. We also give the similarity between the uttered emotion and the *mean* of the emotions contained in the granularity adapted class, by computing the mean of the adapted class's emotions in Euclidean space, and then computing the distance between the single utterance and this mean. Finally, the average for all similarity computations is taken, for each of the experimental cases. These averages give a single score for the baseline case, upper bound (most similar), lower bound (least similar) and mean case, and are presented in columns six, seven, eight and nine, respectively.

6 Discussion and Results

To assist the reader, we have sorted the subject data in order of most number of adapted classes to least number. One of the first things that can be noticed is that there is a strong inverse correlation ($r = -0.72$) between the number of adapted classes extracted for each subject and the accuracy of the adapted classifier (as expected). The lower the number of adapted classes, the easier it is to differentiate between them. However, this is not a strict rule - a case in point is the first two users, who's adapted classifiers contain the same number of adapted classes, but different classification accuracies. In both cases, the total number of adapted classes is 11, where 10 classes are assigned a single emotion, and the 11^{th} class is assigned two emotions. In the classifier with the highest accuracy, we noticed that in adapted class with two emotions, that these emotions were *joy* and *amusement*, whereas in the lower accuracy classifier, the emotions were instead *joy* and *pleasure*. While both configurations make perfect sense, in terms of locality of emotions, one configuration outperforms the other - the accuracy is dependent on what emotions are clustered together. We notice in general, for all users where the number of adapted classes is less than six, that the classification accuracy hovers around 43 %. In general however, high-granularity subjects ended up with fewer of these classes, while for low-granularity subjects, the number of classes was not substantially reduced compared to the baseline. For these low granularity users, we saw that there is very little to be gained by

employing granularity adaptation, and that in fact for one of the subjects (2) the maximum attainable similarity could not be improved. From a similarity-of-emotions perspective, we see that adapted classification has the ability to include more similar items for presentation to subjects. This can be seen in the last row of the table, where a best attainable similarity mean of 0.75 was obtained, as opposed to 1.05 for the baseline (lower is better). This is a potential improvement of 28.57 %. Treating the null hypothesis that the difference between the baseline and adapted case comes from a distribution of zero median, a sign rank test rejects the null hypothesis at the $z = -3.18$, $p < 0.01$ level. We also compute the effect size $r = \frac{Z}{\sqrt{N}} = -0.58$ which corresponds to a large effect size. This confirms our hypothesis, that applying granularity adapted classification can result in more similar emotions being included than for the baseline case.

Looking at the lower bound, we find it to be 1.96 and the mean to be 1.38. In none of the subject cases did either the lower bound or mean improve on the baseline. This reason for this is simply that, as the size of the granularity adapted class, and corresponding region grows, not only will more similar items be available[5], but also more dissimilar items. Even if the adapted class is a good estimate, if it contains more than one emotion, it cannot perform similarity-wise better than a correctly-predicted emotion from the baseline. To show this, in another experiment, we compared results for the adapted classifier and baselines classifiers for just the utterances incorrectly predicted in the baseline case. For just the incorrectly predicted utterances, we found a main baseline similarity of 1.84, and a mean region similarity of 1.68. This indicates that adapted classes containing dissimilar items still perform better than the baseline case for incorrectly predicted items (the main weakness of the traditional classifier).

This large number of dissimilar items poses an interesting problem when considering such a system for recommendation purposes, since while users are indeed given more similar items, they are also given more dissimilar items. This is also why simply reducing the number of classes, without regard for the granularity of users, is not a very good idea. While we expect a large number of dissimilar items to only be present for high-granularity users, this creates noise and would typically mean having to select amongst a large list of items. If items are uniformly spaced, the list would be proportional to the region size. Limiting the number of dissimilar items will both decrease the size of the list and help in the selection process. We propose a simple approach, that while having the undesirable effect of marginally increasing the upper bound, provides a dramatic improvement in both the lower bound case and mean case. Without any further information, we can use to our advantage the fact that each user has two classifiers to their disposal. In the adapted classifier case, when an adapted class for a certain utterance is proposed, there is a certain likelihood that one of the emotions in that class will match the

[5]The extreme case would be a single region occupying all of VA space containing all 12 emotions with a mean best attainable similarity of 0.00 %.

Subject Id	Replace if match Upper bound similarity: Adapted	Replace if match Lower bound similarity: Adapted	Replace if match Mean similarity Adapted	Replace if match Proportional Upper bound similarity: Adapted	Replace if match Proportional Lower bound similarity: Adapted	Replace if match Proportional Mean similarity Adapted
1	1.03	1.04	1.04	1.02	1.08	1.05
2	1.09	1.10	1.10	1.08	1.18	1.13
3	1.05	1.07	1.06	1.04	1.13	1.09
4	0.98	1.33	1.16	0.90	1.55	1.24
5	0.95	1.43	1.23	0.89	1.69	1.33
6	0.95	1.43	1.21	0.87	1.70	1.32
7	0.88	1.23	1.06	0.83	1.45	1.15
8	0.87	1.26	1.08	0.82	1.54	1.20
9	0.76	1.56	1.19	0.60	2.15	1.40
10	0.91	1.51	1.21	0.75	2.00	1.36
11	1.08	1.30	1.20	0.95	1.70	1.33
12	0.96	1.39	1.18	0.65	2.11	1.41
13	0.92	1.46	1.21	0.51	2.25	1.48
14	0.96	1.25	1.12	0.80	1.68	1.27
15	0.98	1.51	1.26	0.74	2.07	1.44
Mean	0.96	1.33	1.15	0.83	1.69	1.28
SD	0.08	0.16	0.07	0.16	0.36	0.13

Table D.3: Results obtained from replacing regions with single emotions from the baseline. Replace if match: best attainable similarity, worst attainable similarity, mean similarity, Proportional Replacement if match: best attainable similarity, worst attainable similarity, mean similarity. Shown for all users including the mean and standard deviation.

single-class baseline prediction for that same utterance (of course we don't know whether this baseline prediction is right or wrong). When we find this to be the case, we replace the larger adapted class with the single emotion class common to both the baseline class and adapted class. In some cases, the baseline emotion is an incorrectly predicted emotion, and will result in a higher dissimilarity overall (undesirable). When the baseline emotion is correctly predicted, the elimination of the remaining emotions from the adapted class that contribute to dissimilarity will result in a lower overall dissimilarity (desired).

The first two columns of Table D.3 show the results of applying this scheme. While the worst attainable similarity decreases from 1.96 down to 1.31 (33.06 %) and the mean similarity decreases from 1.38 down to 1.15 (17.09 %), the highest attainable similarity increases from 0.75 to 0.97 (29.33 %, $z = -2.90$, $p < 0.01$, $r = -0.53$).

Emotion	Probability	Emotion	Probability
Amusement	25.00	Joy	35.00
Pride	25.00	Desperation	70.00
Hot Anger	40.00	Fear	40.00
Anxiety	25.00	Cold Anger	63.16
Sadness	35.00	Interest	60.00
Pleasure	5.00	Relief	95.00

Table D.4: Accuracy of the 12 baseline emotions taken from the validation set.

While we certainly can improve on the lower bound and mean similarity, we do

this at the expense of reducing the number of similar items. If we allow knowledge of the certainty of detecting the different emotions, such as probability estimates taken from a validation set, as shown in Table D.4, which, as expected, is highly correlated with the test results ($r = 0.91$), we can use this knowledge to adjust our approach to consider the replacement of an adapted class with a baseline class, just as in the previous case. However, for the given emotion class, we only do this for the percentage of utterances that equals the probability for that emotion. In other words, the higher the probability of detecting an emotion, the higher the confidence that we can replace the entire adapted class with the single, common emotion. Using such validation set probabilities, and by limiting the replacement of adapted classes to only *probably* correct baseline emotion classes, we end up with an acceptable compromise - as shown in the last two columns of Table D.3, we end up with an adapted class containing a good number of similar emotions (best similarity 0.83, $z = -3.12$, $p < 0.01$, $r = -0.57$), with still a fair reduction in the number of dissimilar emotions (worst similarity of 1.69 and mean similarity of 1.28).

References

[1] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[2] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[3] T. Chambel, E. Oliveira, and P. Martins, "Being happy, healthy and whole watching movies that affect our emotions," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 35–45.

[4] M. Tkalcic, A. Kosir, and J. Tasic, "Affective recommender systems: the role of emotions in recommender systems," in *Proceedings of the 5th ACM conference on recommender systems*, 2011, pp. 9–13.

[5] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143–154, 2005.

[6] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 10, pp. 1175–1191, 2001.

[7] E. Oliveira, P. Martins, and T. Chambel, "Accessing movies based on emotional impact," *Multimedia systems*, vol. 19, no. 6, pp. 559–576, 2013.

[8] A. Milton and S. Tamil Selvi, "Class-specific multiple classifiers scheme to recognize emotions from speech signals," *Computer Speech & Language*, 2013.

[9] R. Xia and Y. Liu, "Using I-vector space model for emotion recognition." in *INTERSPEECH*, 2012, pp. 2230–2233.

[10] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception." *Emotion*, vol. 12, no. 5, p. 1161, 2012.

[11] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 148–152.

[12] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. J. Russell, and M. Wong, ""You Stupid Tin Box"-Children interacting with the AIBO robot: A cross-linguistic emotional speech corpus." in *LREC*, 2004, pp. 171–174.

[13] L. F. Barrett, "Feelings or words? understanding the content in self-report ratings of experienced emotion." *Journal of personality and social psychology*, vol. 87, no. 2, p. 266, 2004.

[14] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[15] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1. IEEE, 2006, pp. 1136–1139.

[16] J. A. Russel, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1170, 1980.

[17] N. A. Remington, L. R. Fabrigar, and P. S. Visser, "Reexamining the circumplex model of affect." *Journal of personality and social psychology*, vol. 79, no. 2, p. 286, 2000.

[18] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, 2005.

[19] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[20] J. Buckow, A. Batliner, R. Huber, H. Niemann, E. Nöth, and V. Warnke, "Detection of prosodic events using acoustic-prosodic features and part-of-speech tags," in *Proceedings of the 5th International Workshop Speech and Computer (SPECOM)*, 2000, pp. 63–66.

[21] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[22] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech." in *Interspeech*, 2005, pp. 1517–1520.

[24] Y. Attabi and P. Dumouchel, "Anchor models for emotion recognition from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 280–290, 2013.

[25] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[26] C. Peter and A. Herbon, "Emotion representation and physiology assignments in digital systems," *Interacting with Computers*, vol. 18, no. 2, pp. 139–170, 2006.

[27] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A dimensional approach to emotion recognition of speech from movies," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 65–68.

[28] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[29] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 566–569.

[30] S. Shepstone, Z.-H. Tan, and S. H. Jensen, "Using audio-derived affective offset to enhance TV recommendation," *Multimedia, IEEE Transactions on*, vol. 16, no. 7, pp. 1999–2014, 2014.

[31] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[32] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[33] G. Gosztolya, R. Busa-Fekete, and L. Tóth, "Detecting autism, emotions and social signals using AdaBoost," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2013, pp. 1–5.

[34] P. J. Lang, M. M. Bradley, B. N. Cuthbert *et al.*, "International affective picture system (IAPS): Instruction manual and affective ratings," *The center for research in psychophysiology, University of Florida*, 1999.

[35] L. A. Feldman, "Valence focus and arousal focus: Individual differences in the structure of affective experience." *Journal of personality and social psychology*, vol. 69, no. 1, p. 153, 1995.

[36] IBM, "IBM SSPS Software, [Online]. available: http://www-01.ibm.com/software/analytics/spss/," 2014.

Paper E

Source-specific Informative Prior for i-Vector Extraction

Sven Ewan Shepstone, Kong Aik Lee, Haizhou Li, Zheng-Hua Tan, and Søren Holdt Jensen

The paper will be publication in the
Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

The layout has been revised.

Abstract

An i-vector is a low-dimensional fixed-length representation of a variable-length speech utterance, and is defined as the posterior mean of a latent variable conditioned on the observed feature sequence of an utterance. The assumption is that the prior for the latent variable is non-informative, since for homogeneous datasets there is no gain in generality in using an informative prior. This work shows that extracting i-vectors for a heterogeneous dataset, containing speech samples recorded from multiple sources, using informative priors instead is applicable, and leads to favorable results. Tests carried out on the NIST 2008 and 2010 Speaker Recognition Evaluation (SRE) dataset show that our proposed method beats three baselines: For the short2-short3 core-task in SRE'08, for the female and male cases, five and six respectively, out of eight common conditions were beaten, and for the core-core task in SRE'10, for both genders, five out of nine common conditions were beaten.

1 Introduction

In the i-vector approach, variable-length speech utterances are mapped into fixed-length low dimensional vectors that reside in the so-called total variability space [1]. The i-vectors capture the *total* variability, which is usually understood to include both speaker and channel variability. The ease of dealing with i-vectors has resulted in a myriad of techniques being proposed to maximize speaker discrimination and reduce channel effects, which include amongst others *within-class covariance normalization* (WCCN) [2], *linear discriminant analysis* (LDA) [3], and *probabilistic LDA* (PLDA) [4].

When i-vectors are extracted from a heterogeneous dataset, as encountered in the recent NIST SREs [5, 6], not only will they capture both speaker and channel variability, but also source variation. If this source variation is not dealt with, it will adversely affect speaker recognition performance [3, 7]. The notion of source variation was introduced in the recent SREs and it is related to the speech acquisition method (e.g., telephone versus microphone channel types) and recording scenario (e.g., telephone conversation versus interview styles). The various combinations of styles and channel types (e.g., interview speech recorded over microphone channel) form relatively homogeneous subsets of the dataset. In this work, the dataset consists of *telephone*, *microphone* (telephone conversation recorded over microphone channel), and *interview* subsets, or sources.

Several proposals consider the issue of source variation within the context of total variability modeling. In [8], the authors address the issue of estimating the inter-speaker scatter matrix given a heterogeneous dataset where most speakers appear only once in any one of the sources. The source variation will be strongly represented and seen as part of the inter-speaker variability and will therefore be

optimized in the resulting LDA transform. Another proposal involves training of a supplementary matrix for the *microphone* subset on top of an already trained total variability matrices on *telephone* data [3]. I-vectors are then extracted from a total variability matrix formed by concatenating the two matrices. PLDA has also been used to further project microphone and telephone factors to a common space [9]. Compensation using heavy-tailed PLDA has also been successful [10]. Finally, a total variability matrix can be trained from a pooled set of the training data. All these schemes require either training of a supplementary matrix or retraining of the total variability matrix.

This work proposes to deal with the source variability by using an informative prior at the i-vector extraction stage. The objective is to use the same total variability matrix to describe the speaker and channel variability across sources of data from a heterogeneous dataset, with the source variation modeled at the priors. Re-training of the total variability matrix is not required, neither in whole or in part. Instead we assume a matrix already trained using abundantly available data. We show how a *source-specific* prior can be used in the i-vector extraction phase to compensate for unwanted source variability. The extracted i-vectors, which now only capture speaker and channel variability, can be processed at the LDA or PLDA stages without needing to carry out any source variation suppression.

This paper is structured as follows: Section 2 reviews the i-vector paradigm and the use of the non-informative prior. Section 3 gives the motivation for using an informative prior when a heterogeneous dataset is concerned. Section 4 presents theory for estimating the source-specific priors and using them effectively in extracting i-vectors. The following two section present the experiments that were carried out and our results, and the final section concludes the paper.

2 The I-vector Paradigm

The total variability model assumes that a speaker- and channel-dependent GMM supervector $\mathbf{m}$ of an utterance [11] is modeled as

$$\mathbf{m} = \mathbf{m_0} + \mathbf{Tw} \tag{E.1}$$

where $\mathbf{m_0}$ is the speaker-independent supervector obtained by concatenating the mean vectors from the UBM. The hidden variable $\mathbf{w}$ weights the columns of the matrix $\mathbf{T}$ to explain the observed deviation from $\mathbf{m_0}$. The matrix $\mathbf{T}$ is defined to have low rank so as to model the subspace where both the speaker and channel variability (hence the name total variability matrix) correlate the most. The training of the total variability matrix follows the same process as that of training an eigenvoice matrix [12, 13]. The major difference is that utterances from the same speakers are treated individually as unrelated sessions [1].

Let $\{\mathbf{o}_1, \mathbf{o}_2, ...\mathbf{o}_T\}$ represent the feature sequence of a given utterance O. The feature vectors are assumed to be drawn from a GMM with its mean supervector as

in (E.1). For each mixture component c of the GMM, the following Baum-Welch statistics are defined:

$$N(c) = \sum_t \gamma_t(c) \tag{E.2}$$

where t extends over all frames of an utterance and $\gamma_t(c)$ is the occupancy of frame $\mathbf{o}_t$ to the c-th Gaussian. We further denote the centered first-order statistics as

$$\widetilde{\mathbf{F}}(c) = \sum_t \gamma_t(c)(\mathbf{o}_t - \mathbf{m_0}(c)) \tag{E.3}$$

Also, let $\mathbf{N}$ represent the diagonal matrix whose diagonal blocks are $N(c) \times \mathbf{I}$ and let $\widetilde{\mathbf{F}}$ represent the supervector obtained by concatenating the $\widetilde{\mathbf{F}}(c)$, where c extends over all mixtures in both cases. In order to extract an i-vector, given an already trained $\mathbf{T}$, we compute the posterior distribution over the latent variable $\mathbf{w}$ conditioned on the observations. Assuming a standard normal prior $\mathbf{w} \sim \mathcal{N}(0, I)$, the posterior distribution is also Gaussian [12], as follows

$$p(\mathbf{w}|O) = \mathcal{N}(\mathbf{L}^{-1} \cdot \mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}},\ \mathbf{L}^{-1}) \tag{E.4}$$

with mean vector

$$\phi = \mathbf{L}^{-1} \cdot \mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}} \tag{E.5}$$

and precision $\mathbf{L} = (\mathbf{I} + \mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{T})$. The i-vector is then given by the mean vector ϕ of the posterior distribution [1]. Similar to that of $\mathbf{N}$, the matrix $\mathbf{\Sigma}$ in (E.4) is constructed by having its diagonal blocks made up by the covariance matrices of the UBM.

The prior over the hidden variable $\mathbf{w}$ is usually taken to be a standard normal distribution. While it is indeed possible to define an informative prior, this prior can always be absorbed to the global mean vector $\mathbf{m_0}$ and the loading matrix $\mathbf{T}$ [13, 14]. This step causes the resulting prior to become non-informative, thereby requiring no alteration to (E.4). As such, there is no compelling reason to use an informative prior at least for the case when the dataset is homogeneous. In the following, we show how informative priors of the form $\mathbf{w} \sim \mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$, where $\mu_\mathbf{p} \neq 0$ and $\mathbf{\Sigma_p} \neq \mathbf{I}$, could be modeled and used for i-vector extraction, and the benefit of doing so when a heterogeneous dataset is concerned. In the NIST series of speaker recognition evaluations (SREs), for instance, the dataset contains "telephone", "interview" or "microphone" speech sources [5, 6].

3 Introducing Informative Priors

An informative prior encodes domain knowledge (i.e., the source variation) by capturing underlying dependencies between the parameters [15]. In this section,

we propose using minimum divergence criterion for estimating source-specific priors from a heterogeneous dataset. We then show how to incorporate the informative prior in the i-vector extraction formula.

3.1 Minimum Divergence Estimation

Consider the case where individual speech sources (e.g., telephone, microphone, or interview in NIST SRE) forms a relatively homogeneous subset and each speech source has I number of utterances. For each utterance we compute the posterior distribution according to (E.4) using the already trained $\mathbf{T}$ matrix. Given the set of posterior distributions, we seek for a Gaussian distribution $\mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma}_{\mathbf{p}})$ that best describes the I posterior distributions. This could be achieved by minimizing the Kullback-Leibler (KL) divergence of the desired distribution $\mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma}_{\mathbf{p}})$ from all the I posteriors $\mathcal{N}(\phi_i, \mathbf{L}_i^{-1})$. As shown in [16], the closed form solution consists of the mean vector

$$\mu_{\mathbf{p}} = \frac{1}{I}\sum_{i=1}^{I}\phi_i \tag{E.6}$$

and the covariance matrix

$$\mathbf{\Sigma}_{\mathbf{p}} = \frac{1}{I}\sum_{i=1}^{I}(\phi_i - \mu_{\mathbf{p}})(\phi_i - \mu_{\mathbf{p}})^{\mathrm{T}} + \frac{1}{I}\sum_{i=1}^{I}\mathbf{L}_i^{-1} \tag{E.7}$$

Notice that the number of utterances I is generally different for each speech source. The central idea here is to use a single $\mathbf{T}$ matrix for all sources of data, where the variability due to the different sources is modeled at the prior. Together, the combination of $\mathbf{T}$ and the source-specific priors better models the variation across sources from the heterogeneous dataset.

Notice that the mean $\mu_{\mathbf{p}}$ of the informative prior is given by the average of all the i-vectors belonging to a target set (recall that an i-vector is given by the mean of the posterior distribution). The deviation of the i-vectors from $\mu_{\mathbf{p}}$ forms the empirical term in the covariance $\mathbf{\Sigma}_{\mathbf{p}}$, while the second term accounts form posterior covariances of the i-vectors.

3.2 Posterior Inference with Informative Prior

We formulate the expression for the posterior distribution for the general case when the informative prior as estimated above is used in place of a non-informative one.

Proposition 1: Consider an informative prior $p(\mathbf{w}) \sim \mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma}_{\mathbf{p}})$ with mean $\mu_{\mathbf{p}}$ and the covariance matrix $\mathbf{\Sigma}_{\mathbf{p}}$. The posterior distribution $p(\mathbf{w}|O)$ is Gaussian with mean

$$\phi = \mathbf{L}^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}} + {\mathbf{\Sigma}_{\mathbf{p}}}^{-1}\mu_{\mathbf{p}}) \tag{E.8}$$

and precision

$$\mathbf{L} = \mathbf{T}^{\mathrm{T}}\mathrm{N}\boldsymbol{\Sigma}^{-1}\mathbf{T} + {\boldsymbol{\Sigma}_{\mathbf{p}}}^{-1} \tag{E.9}$$

Note that by setting $\mu_{\mathbf{p}} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{\mathbf{p}} = \mathbf{I}$, the posterior mean ϕ (i.e., the i-vector) and precision $\mathbf{L}$ reduce to the standard form of i-vector extraction with a non-informative prior as in (E.4).

Proof. Assume that we have the parameter set $(\mathbf{T}, \boldsymbol{\Sigma})$, the hidden variable $\mathbf{w}$ and the observation O. From Lemma 1 in [12] we know that the log likelihood of O given $\mathbf{w}$ and the parameters $(\mathbf{T}, \boldsymbol{\Sigma})$ can be expressed as the sum of two terms:

$$\log p_{\mathbf{T},\boldsymbol{\Sigma}}(O|\mathbf{w}) = G_{\mathbf{T}} + H_{\mathbf{T},\boldsymbol{\Sigma}} \tag{E.10}$$

where $G_{\mathbf{T}}$ is defined by (3) in [12], and $H_{\mathbf{T},\boldsymbol{\Sigma}}$ is defined as

$$H_{\mathbf{T},\boldsymbol{\Sigma}} = \mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{F}} - \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{N}\boldsymbol{\Sigma}^{-1}\mathbf{T}\mathbf{w} \tag{E.11}$$

Since $G_{\mathbf{T}}$ does not depend on $\mathbf{w}$, this term is not considered further. Given the mean $\mu_{\mathbf{p}}$ and covariance ${\boldsymbol{\Sigma}_{\mathbf{p}}}^{-1}$, we express the prior as:

$$p(\mathbf{w}) \propto exp(-\frac{1}{2}(\mathbf{w} - \mu_{\mathbf{p}})^{\mathrm{T}}{\boldsymbol{\Sigma}_{\mathbf{p}}}^{-1}(\mathbf{w} - \mu_{\mathbf{p}})) \tag{E.12}$$

The posterior distribution of $\mathbf{w}$ given O could be obtained by taking the product of (E.11) and (E.12), as follows:

$$\begin{aligned} p(\mathbf{w}|O) &\propto exp(\mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{F}\mathbf{t} - \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{N}\boldsymbol{\Sigma}^{-1}\mathbf{T}\mathbf{w} - \\ &\qquad \frac{1}{2}(\mathbf{w} - \mu_{\mathbf{p}})^{\mathrm{T}}{\boldsymbol{\Sigma}_{\mathbf{p}}}^{-1}(\mathbf{w} - \mu_{\mathbf{p}})) \\ &\propto exp(-\frac{1}{2}(\mathbf{w} - \phi)^{\mathrm{T}}\mathbf{L}(\mathbf{w} - \phi)) \end{aligned} \tag{E.13}$$

with ϕ and $\mathbf{L}$ in the form as stated above. □

4 Prior-compensated i-vector Extraction

In the Bayesian sense, an informative prior increases the prior belief of the location and dispersion of each source in a heterogeneous dataset. We note that a different spread is observed for each source in the i-vector space, as was also reported in a previous study [7]. In the case of cross-source trials, the test i-vectors belonging to one source and target i-vector belonging to another can no longer be assumed to lie close to one another, even when representing the same speaker. The implication

of applying (E.8) directly would intensify the difference across speech sources, resulting in poorer performance.

We propose to compensate for the differences across speech sources (e.g., telephone versus microphone) by applying the prior mean and covariance at separate stages in the i-vector extraction phase. More specifically, we project the prior mean to the acoustic space, while the covariance remains intact as part of the prior. The operation of separating the prior mean and covariance is based on the equality of marginalization which we shall now demonstrate.

Proposition 2: Let $\Pi(c)$ be the marginal distribution for Gaussian c obtained by modeling $\mathbf{m} = \mathbf{m_0} + \mathbf{Tw}$ with the prior $\mathbf{w} \sim \mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$. For this source, the same marginalization $\Pi(c)$ can be realized by modeling $\mathbf{m} = \mathbf{m_0} + \mathbf{Tw} + \mathbf{T}\mu_\mathbf{p}$ with the prior $\mathbf{w} \sim \mathcal{N}(0, \mathbf{\Sigma_p})$. This gives the following equality:

$$\begin{aligned}\Pi(c) &= \int \mathcal{N}(O|\mathbf{m_0}(c) + \mathbf{T}_c\mathbf{w}, \Sigma_0)\mathcal{N}(\mathbf{w}|\mu_\mathbf{p}, \mathbf{\Sigma_p})d\mathbf{w} \\ &= \int \mathcal{N}(O|\mathbf{m_0}(c) + \mathbf{T}_c\mu_\mathbf{p} + \mathbf{T}_c\mathbf{w}, \Sigma_0)\mathcal{N}(\mathbf{w}|0, \mathbf{\Sigma_p})d\mathbf{w}\end{aligned} \tag{E.14}$$

The proof of the proposition is given in the appendix.

Comparing the first and second rows of (F.11), the prior mean $\mu_\mathbf{p}$ is brought forward to the conditional density, which describes the acoustic observation O. By doing so, the projection $\mathbf{T}_c\mu_\mathbf{p}$ of the prior mean imposes a shift on the global mean vector $\mathbf{m_0}(c)$. This also gives rise to prior distributions with a common mode at the origin (i.e., zero mean) but different dispersions $\mathbf{\Sigma_p}$ for individual sources. Algorithmically, the projection $\mathbf{T}_c\mu_\mathbf{p}$ is applied on the observation by re-centering the first order statistics $\widetilde{\mathbf{F}}(c)$, as follows

$$\begin{aligned}\widetilde{\widetilde{\mathbf{F}}}(c) &= \sum_t \gamma_t(c)(\mathbf{o}_t - \mathbf{m_0}(c) - \mathbf{T}_c\mu_\mathbf{p}) \\ &= \widetilde{\mathbf{F}}(c) - N(c)\mathbf{T}_c\mu_\mathbf{p}\end{aligned} \tag{E.15}$$

In a sense, the re-centering brings heterogeneous sources to a common mode at the origin of the total variability space and allows the priors to differ only with regard to one anothers' covariance.

The proposed prior-compensated i-vector extraction can be summarized into the following steps:

1. Start out with an already trained $\mathbf{T}$ matrix. For each source, extract an informative prior $\mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$ using the minimum divergence estimation as described in Section 3.1.

2. Re-center the first order statistics $\widetilde{\mathbf{F}}$ around the relevant source-specific mean to give $\widetilde{\widetilde{\mathbf{F}}}$, as in (E.15).

3. Extract i-vectors, by matching the now zero-mean informative prior $\mathcal{N}(0, \mathbf{\Sigma_p})$ for each source to the relevant re-centered first-order statistics:

$$\begin{aligned}\phi =& \mathbf{L}^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\widetilde{\mathbf{F}} - \mathbf{NT}\mu_{\mathbf{p}})) \\ =& \mathbf{L}^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\widetilde{\mathbf{F}}})\end{aligned} \tag{E.16}$$

where the precision $\mathbf{L}$ is as given in (E.9).

5 Experiments

5.1 Datasets and System Setup

Our experiments were carried out on the short2-short3 core-task of SRE'08 [5] and the core-core task of SRE'10 [6]. For all experiments, a gender dependent setup was used. The features used for training the 512-Gaussian UBMs were 57-dimensional MFCCs (including the first and second derivatives). The first order statistics used for training each total variability matrix were centered and whitened [17]. For all experimental setups, a total variability matrix was trained with non-informative priors being used in the E-step.

We compare four individual experimental setups in this work, of which three are reference systems and one is the proposed system. In the *telephone only* setup, a 600 dimensional $\mathbf{T}$ matrix was trained using only the telephone data. In the *pooled* system, a 600 dimensional $\mathbf{T}$ matrix was trained using pooled telephone and microphone data. In the *cascade* system, a 400 dimensional $\mathbf{T}$ matrix was trained using the telephone data, and a 200 dimensional $\mathbf{T}$ matrix was trained using microphone data [3]. The telephone data used to train these systems was taken from SRE'04, 05 and 06. The microphone data was taken from SRE'05, 06 and MIXER 5. The same dataset was used to derive the informative priors.

In the *2-prior* system, the already trained *pooled* $\mathbf{T}$ matrix was used as the starting point. Using minimum divergence estimation (Section 3.1), we trained one prior for the telephone subset and another prior for microphone and interview subsets. We chose to use only one prior for both microphone and interview since there was not enough interview data to reliably estimate the interview prior. I-vectors were extracted by performing re-centering of the first-order statistics using the prior's mean, followed by computation of the posteriors using the prior's informative covariance. LDA was used to bring the dimension of the 600-dimensional i-vectors down to 400. After carrying out length normalization, PLDA was used to model the channel variability. For the PLDA model, a separate 200 dimensional telephone matrix and 50 dimensional microphone matrix were trained, in a decoupled manner, similar to the setup in [18].

	CC1: int-int		CC2: int-int		CC3: int-int		CC4: int-tel		CC5: tel-mic		CC6: tel-tel		CC7: tel-tel		CC8: tel-tel	
EER	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
Telephone only	3.51	2.84	1.50	0.32	3.61	2.97	5.69	4.07	6.65	4.17	5.85	4.67	2.73	2.32	3.24	1.43
Pooled	3.22	2.54	1.28	0.33	3.29	2.64	4.65	3.89	5.62	3.05	5.86	4.15	2.84	1.60	3.32	1.04
Cascade	3.17	3.01	1.25	0.41	3.26	3.22	5.38	4.27	6.10	4.12	5.86	4.06	2.98	1.66	3.81	1.32
2-prior	**2.34**	**1.95**	1.32	**0.32**	**2.39**	**2.04**	**4.32**	3.91	**5.37**	3.21	**5.79**	**3.84**	2.87	**1.39**	3.27	**0.90**

Table E.1: SRE'08 Performance comparison for the sub-task short2-short3. Left: FEMALE Trials, Right: MALE Trials

	CC1: int-int-same-mic		CC2: int-int-diff-mic		CC3: int-tel		CC4: int-mic		CC5: nve-nve-diff-tel		CC6: nve-hve-diff-tel		CC7: nve-hve-mic		CC8: nve-lve-diff-tel		CC9: nve-lve-mic	
EER	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
Telephone only	3.06	2.02	5.65	3.45	4.21	3.55	3.96	2.72	3.59	3.47	8.09	4.64	8.49	4.91	2.01	1.14	2.46	1.54
Pooled	3.16	2.22	5.13	3.14	3.34	2.82	3.78	2.54	3.00	2.60	7.13	4.01	7.98	4.95	1.66	1.54	2.55	1.34
Cascade	3.12	2.29	5.60	3.29	4.01	2.62	4.04	2.87	3.41	3.13	7.10	4.33	8.19	5.25	1.83	1.68	3.08	1.61
2-prior	**2.43**	**1.67**	**4.44**	**2.25**	3.87	3.19	**3.33**	**2.22**	**3.00**	2.89	7.11	4.13	**7.49**	**4.16**	**1.59**	1.56	2.48	**1.15**

Table E.2: SRE'10 Performance comparison for the sub-task core-core. Left: FEMALE Trials, Right: MALE Trials

5.2 Results

We present results for the four systems, for both male and female trials. For all results, we used Equal Error Rate (EER). For the SRE'08 results, shown in Table E.2 for both male and female trials, a substantial improvement was seen in sub-tasks 1 and 3, corresponding to the *int-int* condition. We could not beat the baseline for sub-task 2, which we believe is due to the smaller number of trials. For the mixed trials, i.e. sub-tasks 4 and 5, source-specific informative priors showed improved robustness against both the telephone-only and cascade cases. For the pooled case however, the results were a lot closer and we did not beat this baseline in all cases. Interestingly, our approach improved on several of the *tel-tel* only conditions, especially in the male case. From these results, it appears that source-specific informative priors offer the greatest strength in enhancing performance trials where the sources of the trail and target match.

We now discuss the SRE'10 results shown in Table E.2. For the single source interview and mic sub-tasks, as given by sub-tasks 1, 2, 7 and 9, we were able to beat all baselines in 3 out of 4 sub-tasks in the female case and all cases in the male case. For telephone only trials, given by sub-tasks 5, 6 and 8, in only one case could all baselines be beaten. We believe the reason for the slightly worse results for SRE'10 is the similarity of the data used to train the $\mathbf{T}$ matrices and subspace PLDA models to that of SRE'08. For the cross-channel conditions, we noted better performance for the int-mic cross channel than for int-tel, strengthening our belief that best performance is gained where source and target trials are better.

6 Conclusion

In this paper, we proposed a novel method of using a single $\mathbf{T}$ matrix to better describe the source variation from a heterogeneous dataset. The gist of our proposal is to compensate for source variation by applying the prior mean and covariance at separate stages in the i-vector extraction. We showed that by using an existing $\mathbf{T}$ matrix, introducing informative priors for each source into the i-vector extraction stage leads to performance gains in 5 out of 8 and 6 out of 8 common conditions for the short2-short3 core-task in SRE'08 for the female and male case, respectively, and 5 out of 9 common conditions for the core-core task in SRE'10, for both the female and male case. The results show that source-specific informative priors offer the greatest strength in enhancing performance trials where the sources of the trail and target are similar, or match.

A Proof of Proposition 2

Proof. We first derive the probability distribution of $p(\mathbf{m})$ where $\mathbf{m} = \mathbf{m_0} + \mathbf{Tw}$ and $\mathbf{w} \sim \mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$. The mean is computed as:

$$\mathrm{E}[\mathbf{m}] = \mathbf{m_0} + \mathbf{T}\mu_\mathbf{p} \tag{E.17}$$

and covariance as:

$$\mathrm{E}[(\mathbf{m} - \mathrm{E}[\mathbf{m}])^2] = \mathbf{T}\mathrm{E}[\mathbf{w}\mathbf{w}^\mathrm{T}]\mathbf{T}^\mathrm{T} - \mathbf{T}\mu_\mathbf{p}{\mu_\mathbf{p}}^\mathrm{T}\mathbf{T}^\mathrm{T} \tag{E.18}$$

Realizing that the covariance of the prior distribution for $P(\mathbf{w})$ is simply $\mathbf{\Sigma_p} = \mathrm{E}[(\mathbf{w} - \mathrm{E}[\mathbf{w}]^2)] = \mathrm{E}[\mathbf{w}\mathbf{w}^\mathrm{T}] - \mu_\mathbf{p}{\mu_\mathbf{p}}^\mathrm{T}$, substituting back into into (F.25) and simplifying, gives:

$$\mathrm{E}[(\mathbf{m} - \mathrm{E}[\mathbf{m}])^2] = \mathbf{T}\mathbf{\Sigma}_p\mathbf{T}^\mathrm{T} \tag{E.19}$$

Note that for the case of the non-informative prior, the mean and covariance are reduced to $\mathbf{m_0}$ and $\mathbf{T}\mathbf{T}^\mathrm{T}$, respectively. In the same vein, we compute the mean for the marginalization modeled by $\mathbf{m} = \mathbf{m_0} + \mathbf{Tw} + \mathbf{T}\mu_\mathbf{p}$ and $\mathbf{w} \sim \mathcal{N}(0, \mathbf{\Sigma_p})$. We find the mean to be

$$\mathrm{E}[\mathbf{m}] \quad = \mathbf{m_0} + \mathbf{T}\mu_\mathbf{p} \tag{E.20}$$

which is identical to the formally derived mean. The covariance is computed as:

$$\mathrm{E}[(\mathbf{m} - \mathrm{E}[\mathbf{m}])^2] = \mathbf{T}\mathrm{E}[\mathbf{w}\mathbf{w}^\mathrm{T}]\mathbf{T}^\mathrm{T} \tag{E.21}$$

Now the covariance $\mathbf{\Sigma}_p = \mathrm{E}[(\mathbf{w} - \mathrm{E}[\mathbf{w}])^2] = \mathrm{E}[\mathbf{w}\mathbf{w}^\mathrm{T}]$, which when substituted back into (F.29), gives:

$$\mathrm{E}[(\mathbf{m} - \mathrm{E}[\mathbf{m}]^2)] = \mathbf{T}\mathbf{\Sigma}_p\mathbf{T}^\mathrm{T} \tag{E.22}$$

which is identical to the formally derived covariance. These will contribute equally to the marginalization $\Pi(c)$ given in (F.11). This concludes the proof. □

References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[2] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition." in *Interspeech*, 2006, pp. 1471–1474.

[3] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech." in *Odyssey*, 2010, p. 6.

[4] Y. Jiang, K. A. Lee, and L. Wang, "PLDA in the i-supervector space for text-independent speaker verification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–13, 2014.

[5] National Institute of Standards and Technology, "The NIST 2008 SRE Evaluation Plan," 2008. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2008/

[6] ——, "The NIST 2010 SRE Evaluation Plan," 2010. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/

[7] M. McLaren and D. Van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5456–5459.

[8] ——, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 755–766, 2012.

[9] N. Dehak, Z. N. Karam, D. A. Reynolds, R. Dehak, W. M. Campbell, and J. R. Glass, "A channel-blind system for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4536–4539.

[10] M. Senoussaoui, P. Kenny, P. Dumouchel, and F. Castaldo, "Well-calibrated heavy tailed bayesian speaker verification for microphone speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4824–4827.

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

[12] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[13] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[14] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[15] R. Raina, A. Y. Ng, and D. Koller, "Constructing informative priors using transfer learning," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 713–720.

[16] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. R. Dai, "Minimum divergence estimation of speaker prior in multi-session PLDA scoring," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4007–4011.

[17] P. Kenny, "A small foot-print i-vector extractor," in *Proc. Odyssey*, 2012.

[18] K. A. Lee, A. Larcher, C. H. You, B. Ma, and H. Li, "Multi-session PLDA scoring of i-vector for partially open-set speaker detection." in *INTERSPEECH*, 2013, pp. 3651–3655.

Paper F

Total Variability Modeling Using Source-specific Priors

Sven Ewan Shepstone, Kong Aik Lee, Haizhou Li, Zheng-Hua Tan,
and Søren Holdt Jensen

The paper has been submitted to the
IEEE Transactions on Audio, Speech and Language Processing.

In peer review
The layout has been revised.

Abstract

In total variability modeling, variable length speech utterances are mapped to fixed low-dimensional i-vectors. Central to computing the total variability matrix and i-vector extraction, is the computation of a posterior distribution for a latent variable conditioned on an observed feature sequence of an utterance. In both cases the prior for the latent variable is assumed to be non-informative, since for homogeneous datasets there is no gain in generality in using an informative prior. This work shows in the heterogeneous case, that using informative priors for computing the posterior, can lead to favorable results. We focus on modeling the priors using minimum divergence criterion or factor analysis techniques. Tests on the NIST 2008 and 2010 Speaker Recognition Evaluation (SRE) dataset show that our proposed method beats three baselines: For i-vector extraction using an already trained matrix, for the short2-short3 task in SRE'08, six out of eight common conditions, for both genders, were improved. For the core-core task in SRE'10, six out of nine female and five out of nine male common conditions were improved. When incorporating prior information into the training of the T matrix itself, the proposed method beats the baselines for six out of eight common conditions, for both genders, for SRE'08, and six and five out of nine conditions, for the male and female case, respectively, for SRE'10. Tests using factor analysis for estimating priors show that two priors do not offer much improvement, but in the case of three separate priors (sparse data), considerable improvements were gained.

1 Introduction

The i-vector feature extraction approach has been the state-of-the-art in speaker recognition in recent years. The i-vectors capture the *total* variability, which may include speaker, channel and source variability. Variable-length speech utterances are mapped into fixed-length low dimensional vectors that reside in the so-called total variability space [1].

While it is possible to work directly with the raw i-vector distribution, the fixed-length of i-vectors has resulted in a number of powerful and well-known channel compensation techniques that deal with unwanted channel variability and hence improve speaker recognition performance. As a good starting point, linear discriminant analysis (LDA) is a non-probabilistic method used to further reduce the dimensionality of i-vectors, which simultaneously maximizes the inter-speaker variability and minimizes the intra-speaker variability [2]. After centering and whitening, the i-vectors are more or less evenly distributed around a hypersphere. An important further refinement commonly carried out is length normalization, which transforms the i-vector distribution to an (almost) Gaussian distribution that is more straightforward to model [3]. Probabilistic LDA is a generative model that

uses a factor-analysis approach to model separately factors that account for the inter-speaker and intra-speaker variation [4, 5]. Many variants of PLDA, in the context of the i-vector approach, have been proposed [6, 7]. Another well-known method is within-class covariance normalization (WCCN), which uses the inverse of the within-class covariance matrix to normalize the linear kernel in an SVM classifier [8]. It is typical in i-vector modeling to use multiple techniques in cascade: for example to ensure the Gaussian assumption for PLDA, it is not uncommon to carry out whitening followed by length normalization before the PLDA stage [7, 9].

This work deals primarily with source variation, and due to the fact that channel variation and source variation both contribute to reducing the ability to discriminate speakers, it is not surprising that the methods proposed to combat channel variation and source variation resemble one another. When i-vectors are extracted from a heterogeneous dataset, as encountered in the recent NIST SREs [10, 11], not only will they capture both speaker and channel variability, but also source variation. If this source variation is not dealt with, it will adversely affect speaker recognition performance [2, 12, 13]. The notion of source variation was introduced in the recent SREs and it is related to the speech acquisition method (e.g., telephone versus microphone channel types) and recording scenario (e.g., telephone conversation versus interview styles). The various combinations of styles and channel types (e.g., interview speech recorded over microphone channel) form a heterogeneous dataset consisting of relatively homogeneous subsets. In this work, the dataset consists of *telephone*, *microphone* (telephone conversation recorded over microphone channel), and *interview* subsets, or sources.

There have been several proposals to address the issue of source variation within the context of total variability modeling. A phenomenon commonly seen in heterogeneous datasets is the fact that not all sources are equally abundant and most speakers appear in only one of the sources. In the context of LDA, the source variation will be strongly represented and seen as part of the inter-speaker variability and will therefore be optimized in the resulting LDA transform. One proposal to address this issue is to determine a suitable inter-speaker scatter matrix [12, 13].

For training of the total variability matrix itself, one of the simplest approaches, albeit rather crude, is to simply pool all the training data into a heterogeneous set without distinguishing between microphone and telephone data. A more structured proposal suggests training a supplementary matrix for the *microphone* subset on top of an already trained total variability matrices on *telephone* data [2]. I-vectors are then extracted from a total variability matrix that is formed by concatenating these two matrices. An interesting observation seen with this approach is that the microphone data resides in the combined space defined by the matrix concatenation, whereas the telephone data only resides in the telephone space. An extension to this work was therefore proposed whereby PLDA is applied to project the telephone and microphone data to the same space [14].

In this work we show how informative priors can be estimated from speech data, and subsequently used in the Bayesian sense to annihilate source variability

in total variability models. In JFA [15, 16] and total variability modeling [1], a non-informative prior is assumed for the speaker, channel and total variability latent variables, since there is no gain in generality in using an informative prior. This assertion holds at least when a homogeneous dataset is concerned. The notion of informative priors to encode domain knowledge is not a new concept and has been used in machine learning applications before [17]. In the context of continuous speech recognition, informative priors have also been used in the case of sparse data to improve generalization of an infinite structured SVM model [18].

The contribution of this paper is to propose and investigate three different strategies for incorporating prior information into different aspects of total variability modeling. The first strategy involves using an already trained total variability matrix to extract i-vectors [19], and comprises two stages. In the first stage, i-vectors from each subset of the data are extracted using the standard non-informative prior, and then all i-vectors are subsequently used to estimate a source-specific prior. The second stage comprises using the source-specific prior in the computation of the posterior to compute a new set of i-vectors. The hypothesis of this strategy is that i-vector extraction using source-specific priors can be used to compensate for unwanted source variability.

Since the ultimate performance will be affected by the initial alignment of the total variability matrix, as a second strategy, we propose to retrain the total variability matrix. Here, we extend the role of the source-specific prior to the computation of the posterior mean and covariance in the E-step, needed for re-estimating the total variability matrix for a given training iteration. For each training iteration, we recompute the source-specific prior and use it to update the total variability matrix. Once the training has completed, we treat this new total variability matrix as the already existing matrix proposed in the first strategy and follow the same approach for extracting i-vectors. The hypothesis of this strategy is that, assuming that i-vectors are extracted according to the first strategy, that performance can be improved by using prior information to improve the initial alignment of the total variability matrix. As the third strategy, noting that performance may be influenced by the manner and accuracy with which the prior is estimated, we propose and investigate the use of factor analysis for estimating the priors. In this approach, both the mean and covariance of the posterior (where the mean corresponds to our i-vector) are taken into account.

This paper is structured as follows: Section 2 reviews the total variability paradigm and the use of non-informative priors. Section 3 motivates the use of informative priors when a heterogeneous dataset is concerned, and how this can be constructed using minimum divergence estimation. We also show how the informative prior leads to a new formulation for the posterior. Section 4 shows how to use informative priors for i-vector extraction and constructing a new total variability matrix. Section 5 presents theory for factor analysis modeling of priors where the data for a prior could be sparse. The following two sections present the experiments that were carried out and our results, and the final section concludes the paper.

2 Total Variability Modeling

The total variability model assumes that a speaker- and channel-dependent GMM supervector $\mathbf{m}$ of an utterance [1] is modeled as

$$\mathbf{m} = \mathbf{m_0} + \mathbf{Tw} \tag{F.1}$$

Here, assuming C mixture components and an F dimensional feature space, $\mathbf{m_0}$ is the $CF \times 1$ speaker-independent supervector obtained by concatenating the mean vectors from the Universal Background Model (UBM) [20]. The matrix $\mathbf{T}$ is defined to have low rank R so as to model the subspace where both the speaker and channel variability covary the most, and the $R \times 1$ hidden variable $\mathbf{w}$ weights the columns of the $CF \times R$ total variability matrix $\mathbf{T}$ to explain the observed deviation from the global mean. The training of the total variability matrix follows the same process as that of training an eigenvoice matrix [15, 21]. The major difference is that utterances from the same speakers are treated individually as unrelated sessions [1], resulting in the matrix containing the eigenvectors with the largest eigenvalues of the total covariance matrix.

Let $\{\mathbf{o}_1, \mathbf{o}_2, ...\mathbf{o}_T\}$ represent the feature sequence of a given utterance O. The feature vectors are assumed to be drawn from a GMM with its mean supervector as in (F.1). For each mixture component c of the GMM, the following Baum-Welch statistics are defined:

$$N(c) = \sum_t \gamma_t(c) \tag{F.2}$$

where t extends over an utterance and $\gamma_t(c)$ is the occupancy of frame $\mathbf{o}_t$ to the c-th Gaussian. We further denote the centered first-order statistics as

$$\widetilde{\mathbf{F}}(c) = \sum_t \gamma_t(c)(\mathbf{o}_t - \mathbf{m_0}(c)) \tag{F.3}$$

Also, let $\mathbf{N}$ represent the $CF \times CF$ (supervector-size) diagonal matrix whose diagonal blocks are $N(c) \times \mathbf{I}$ and let $\widetilde{\mathbf{F}}$ represent the $CF \times 1$ supervector obtained by concatenating the $\widetilde{\mathbf{F}}(c)$, where c extends over all mixtures in both cases. The training of the $\mathbf{T}$ matrix goes according to the well-known Expectation Maximization (EM) algorithm [21, 22]. The most important computation is the E-step, where for the given sequences of observations, the posterior distributions $p(\mathbf{w}|O)$ are determined for the latent variables. Assuming a standard normal prior $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$, the posterior distribution is also Gaussian [15, 21], and is given as follows

$$p(\mathbf{w}|O) = \mathcal{N}(\mathbf{L}^{-1} \cdot \mathbf{T}^{\mathrm{T}} \mathbf{\Sigma}^{-1} \widetilde{\mathbf{F}},\ \mathbf{L}^{-1}) \tag{F.4}$$

with mean vector

$$\phi = \mathbf{L}^{-1} \cdot \mathbf{T}^{\mathrm{T}} \mathbf{\Sigma}^{-1} \widetilde{\mathbf{F}} \tag{F.5}$$

and precision matrix

$$\mathbf{L} = (\mathbf{I} + \mathbf{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}\mathbf{T}) \tag{F.6}$$

Here, the superscript $^{\mathrm{T}}$ denotes matrix transposition. Similar to that of $\mathbf{N}$, the matrix $\boldsymbol{\Sigma}$ in (F.4) is constructed by having its diagonal blocks made up by the covariance matrices of the UBM. In the M-step, the value for $\mathbf{T}$ is updated by solving a set of simultaneous equations [21]. These EM steps are repeated until convergence. Alg. F.1 lists the details of the EM Step.

Algorithm F.1: Expectation-Maximization (EM) algorithm steps used for each iterative update of the $\mathbf{T}$ matrix given I training utterances [21].

input : $\mathbf{T}, \mathbf{N}, \widetilde{\mathbf{F}}$
output: $\mathbf{T}$
begin
  `// Reset accumulators`
  $\mathbf{A} = \mathbf{0}$
  $\mathbf{C} = \mathbf{0}$
  `// Expectation Step - Compute the posterior` $p(\mathbf{w}\,|\,O) = \mathcal{N}(\phi,\ \mathbf{L}^{-1})$
  for $i = 1$ **to** I **do**
    $\mathbf{L}_i = (\mathbf{I} + \mathbf{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}_i\mathbf{T})$
    $\phi_i = \mathbf{L}_i^{-1} \cdot \mathbf{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{F}}_i$
    $E[\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}] = \phi_i{\phi_i}^{\mathrm{T}} + \mathbf{L}_i$
    `// Accumulate Statistics`
    $\mathbf{A} = \mathbf{A} + \mathbf{N}_i.E[\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}]$
    $\mathbf{C} = \mathbf{C} + \widetilde{\mathbf{F}}_i.\phi_i$
  end
  `// Maximization Step - Solve the set of simultaneous equations:` $\sum_i \mathbf{N}_i\mathbf{T}E[\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}] = \sum_i \mathbf{F}_i\phi_i$
  $\mathbf{T} = \mathbf{C}.\mathbf{A}^{-1}$ `// Solves for` $\mathbf{T}$
end

Extracting an i-vector ϕ is identical to computing the posterior distribution above, where only the posterior mean is retained [1]. Later on, we shall see how this identical process for i-vector extraction and training of the $\mathbf{T}$ matrix allows us to introduce an informative prior at separate stages of the total variability model. In future, when referring to the posterior inference of the hidden variables, the reader should have both simple i-vector extraction, as well as posterior covariance computation (needed when training the $\mathbf{T}$ matrix) in mind.

The prior over the hidden variable $\mathbf{w}$ is usually taken to be a standard normal distribution. While it is indeed possible to define an informative prior, this prior can always be absorbed to the global mean vector $\mathbf{m_0}$ and the loading matrix $\mathbf{T}$ [15, 23]. This step causes the resulting prior to become non-informative, thereby requiring no alteration to (F.4). As such, there is no compelling reason to use an informative prior, at least for the case when the dataset is homogeneous. In fact, this is a common step used in the training of the $\mathbf{T}$ matrix, known as the minimum divergence re-estimation step [15, 16, 24]. It leads to faster convergence and can be carried out once or several times.

3 Prior Modeling

An informative prior encodes domain knowledge (i.e., the source variation) by capturing the underlying dependencies between the parameters [17]. In this section, we show how informative priors of the form $\mathbf{w} \sim \mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$, where $\mu_\mathbf{p} \neq 0$ and $\mathbf{\Sigma_p} \neq \mathrm{I}$, can be modeled and used for computing the posterior distribution. Mathematically, this posterior distribution will have the highest density in regions that agree with both the observed data as well as the prior [5]. Informative priors are of particular benefit when a heterogeneous dataset is concerned. In the NIST series of speaker recognition evaluations (SREs), for instance, the dataset contains "*telephone*", "*interview*" or "*microphone*" speech sources [10, 11]. We propose to estimate the hyper parameters using minimum divergence criterion based on source data from a heterogeneous dataset. A variant of this has been used before in the context of adapting PLDA models [25]. We then show how to incorporate the informative prior in the posterior computation.

3.1 Introducing Informative Priors

Consider the case where individual speech sources (e.g., telephone, microphone, or interview in NIST SRE) forms a relatively homogeneous subset and each speech source has I number of utterances. For each utterance, we compute the posterior distribution according to (F.4) using a given $\mathbf{T}$ matrix. Later we shall see that the precise state of $\mathbf{T}$ (either converged or still undergoing training) that is used depends on the intended use of the prior. Given the set of posterior distributions, we seek for a Gaussian distribution $\mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$ that best describes the I posterior distributions. This can be achieved by minimizing the Kullback-Leibler (KL) divergence of the desired distribution $\mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$ from all the I posteriors $\mathcal{N}(\phi_i, \mathbf{L}_i^{-1})$. As shown in [26], the closed form solution is as follows, consisting of the mean vector

$$\mu_\mathbf{p} = \frac{1}{I} \sum_{i=1}^{I} \phi_i \tag{F.7}$$

and the covariance matrix

$$\mathbf{\Sigma_p} = \frac{1}{I}\sum_{i=1}^{I}(\phi_i - \mu_\mathbf{p})(\phi_i - \mu_\mathbf{p})^{\mathrm{T}} + \frac{1}{I}\sum_{i=1}^{I}\mathbf{L}_i^{-1} \tag{F.8}$$

Notice that the mean $\mu_\mathbf{p}$ of the informative prior is given by the average of all the i-vectors belonging to a target set (recall that an i-vector is given by the mean of the posterior distribution). The deviation of the i-vectors from $\mu_\mathbf{p}$ forms the empirical term in the covariance $\mathbf{\Sigma_p}$, while the second term accounts for the posterior covariances of the i-vectors.

In the above formulation, the number of utterances I could be different for each speech source. Notice also the central idea here is to use a single $\mathbf{T}$ matrix for all sources of data, where the variability due to different sources is modeled at the prior using (F.7) and (F.8). Together, the combination of $\mathbf{T}$ and the source-specific priors better models the variation across sources from the heterogeneous dataset. Both the initial alignment of data in $\mathbf{T}$ as well as the accuracy and strength of the prior will influence the degree at which the source variation can be modeled.

3.2 Posterior Inference with Informative Prior

We now formulate the expression for the posterior distribution for the general case when the informative prior as estimated above is used in place of a non-informative one.

Proposition 1: Consider an informative prior $p(\mathbf{w}) \sim \mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$ with mean $\mu_\mathbf{p}$ and the covariance matrix $\mathbf{\Sigma_p}$. The posterior distribution $p(\mathbf{w}|O)$ is Gaussian with mean

$$\phi = \mathbf{L}^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}} + \mathbf{\Sigma_p}^{-1}\mu_\mathbf{p}) \tag{F.9}$$

and precision matrix

$$\mathbf{L} = \mathbf{T}^{\mathrm{T}}\mathbf{N}\mathbf{\Sigma}^{-1}\mathbf{T} + \mathbf{\Sigma_p}^{-1} \tag{F.10}$$

Note that by setting $\mu_\mathbf{p} = \mathbf{0}$ and $\mathbf{\Sigma_p} = \mathbf{I}$, the posterior mean and precision $\mathbf{L}$ reduce to the standard form in (F.4). The proof of the proposition is in the appendix.

4 Total Variability Modeling Using Multiple Prior

In the Bayesian perspective, an informative prior increases the prior belief of the location and dispersion of each source in a heterogeneous dataset. We note that a different spread is observed for each source in the i-vector space. In the case of cross-source trials, the test i-vectors belonging to one source and target i-vector belonging to another can no longer be assumed to lie close to one another, even when

representing the same speaker. The implication of applying (F.9) directly would intensify the difference across speech sources, resulting in poorer performance. We shall demonstrate further in this section.

4.1 Prior-compensated i-vector Extraction

We propose to compensate for the differences across speech sources (e.g., telephone versus microphone) by applying the prior mean and covariance at separate stages in the i-vector extraction. More specifically, we propagate the prior mean to the acoustic space, while the covariance remains intact as part of the prior. The operation of segregating the prior mean and covariance is based on the equality of marginalization which we shall now demonstrate.

Proposition 2: Let $\Pi(c)$ be the marginal distribution for Gaussian c obtained by modeling $\mathbf{m} = \mathbf{m_0} + \mathbf{T}\mathbf{w}$ with the prior $\mathbf{w} \sim \mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$. The same marginalization $\Pi(c)$ can be realized by modeling $\mathbf{m} = \mathbf{m_0} + \mathbf{T}\mathbf{w} + \mathbf{T}\mu_\mathbf{p}$ with the prior $\mathbf{w} \sim \mathcal{N}(0, \mathbf{\Sigma_p})$. This gives the following equality:

$$\begin{aligned}\Pi(c) &= \int \mathcal{N}(O|\mathbf{m_0}(c) + \mathbf{T}_c\mathbf{w}, \Sigma_0)\mathcal{N}(\mathbf{w}|\mu_\mathbf{p}, \mathbf{\Sigma_p})d\mathbf{w} \\ &= \int \mathcal{N}(O|\mathbf{m_0}(c) + \mathbf{T}_c\mu_\mathbf{p} + \mathbf{T}_c\mathbf{w}, \Sigma_0)\mathcal{N}(\mathbf{w}|0, \mathbf{\Sigma_p})d\mathbf{w}\end{aligned} \tag{F.11}$$

The proof of the proposition is in the appendix.

Comparing the first and second rows of (F.11), the prior mean $\mu_\mathbf{p}$ is brought forward to the conditional density, which describes the acoustic observation O. By doing so, the term $\mathbf{T}_c\mu_\mathbf{p}$ imposes a shift on the global mean vector $\mathbf{m_0}(c)$. This also gives rise to prior distributions with a common mode at the origin (i.e., zero mean) but different dispersions $\mathbf{\Sigma_p}$ for individual sources. Algorithmically, the projection $\mathbf{T}_c\mu_\mathbf{p}$ is applied on the observation by re-centering the first-order statistics $\widetilde{\mathbf{F}}(c)$, as follows

$$\begin{aligned}\widetilde{\widetilde{\mathbf{F}}}(c) &= \sum_t \gamma_t(c)(\mathbf{o}_t - \mathbf{m_0}(c) - \mathbf{T}_c\mu_\mathbf{p}) \\ &= \widetilde{\mathbf{F}}(c) - N(c)\mathbf{T}_c\mu_\mathbf{p}\end{aligned} \tag{F.12}$$

This re-centering process brings heterogeneous sources to a common mode at the origin of the total variability space and allows the priors to differ only with regard to one anothers' covariance.

The proposed prior-compensated i-vector extraction can be summarized into the following steps:

1. Start out with an already trained $\mathbf{T}$ matrix.
2. For each source, extract an informative prior $\mathcal{N}(\mu_\mathbf{p}, \mathbf{\Sigma_p})$ using the minimum divergence estimation as described in Section 3.1.

3. Re-center the first-order statistics $\widetilde{\mathbf{F}}$ around the relevant source-specific mean to give $\widetilde{\widetilde{\mathbf{F}}}$, as in (F.12).

4. Extract i-vectors, by matching the now zero-mean informative prior $\mathcal{N}(0, \mathbf{\Sigma_p})$ for each source to the relevant re-centered first-order statistics:

$$\begin{aligned}\phi =& \mathbf{L}^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\widetilde{\mathbf{F}} - \mathbf{NT}\mu_{\mathbf{p}})) \\ =& \mathbf{L}^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\widetilde{\mathbf{F}}})\end{aligned} \tag{F.13}$$

where the precision $\mathbf{L}$ is as given in (F.10).

In the above process, the prior mean is projected to the acoustic space and applied on the observed feature vectors ((F.12) and (F.13)). The end result is that i-vectors extracted from different sources are being pulled toward a common mode at the origin. In retrospect, the prior covariance could have been used in a similar manner. Nevertheless, this is not done so for the following reasons: Propagating the prior covariance to the acoustic space will cause the $\mathbf{T}$ matrix to scale and rotate differently for each source. As a result, the i-vectors extracted from different sources will lie in vector spaces that do not correspond to one another. Alg. F.2 lists the details of the prior-compensated i-vector extraction steps.

Algorithm F.2: Prior compensated i-vector extraction steps

input : $\mathbf{T}, \mathbf{N}, \widetilde{\mathbf{F}}$
output: ϕ
begin
`// For each source, determine` $\mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma_p})$
$\mu_{\mathbf{p}} = \frac{1}{I}\sum_{i=1}^{I}\phi_i$
$\mathbf{\Sigma_p} = \frac{1}{I}\sum_{i=1}^{I}(\phi_i - \mu_{\mathbf{p}})(\phi_i - \mu_{\mathbf{p}})^{\mathrm{T}} + \frac{1}{I}\sum_{i=1}^{I}\mathbf{L}_i^{-1}$
`// Re-center first order statistics`
$\widetilde{\widetilde{\mathbf{F}}}(c) = \widetilde{\mathbf{F}}(c) - N(c)\mathbf{T}_c\mu_{\mathbf{p}}$
`// Extract i-vector`
$\phi = \mathbf{L}^{-1}(\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\widetilde{\mathbf{F}}})$
end

Figs. F.1 and F.2 show scatter plots of i-vectors extracted from utterances for all three sources, to show the effect of applying source-informative priors. In Fig. F.1, a non-informative prior is used, and here it is evident that there is a different spread in the i-vector space for each source. In Fig. F.2, one informative prior is used for the telephone data and another informative prior is used for the

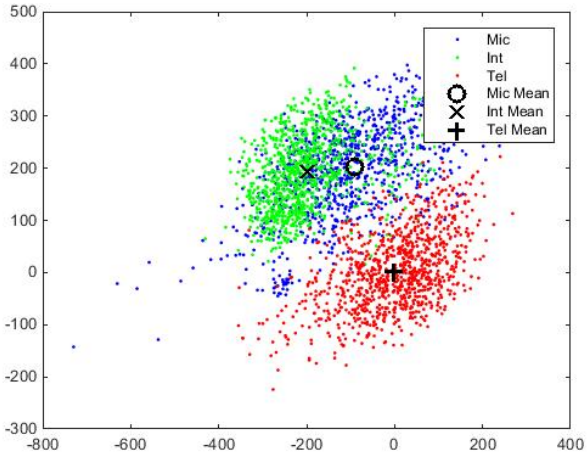


Fig. F.1: Distribution of i-vectors for three sources projected onto 2-dimensional space defined by the first two principal axes of PCA. Non-informative case.

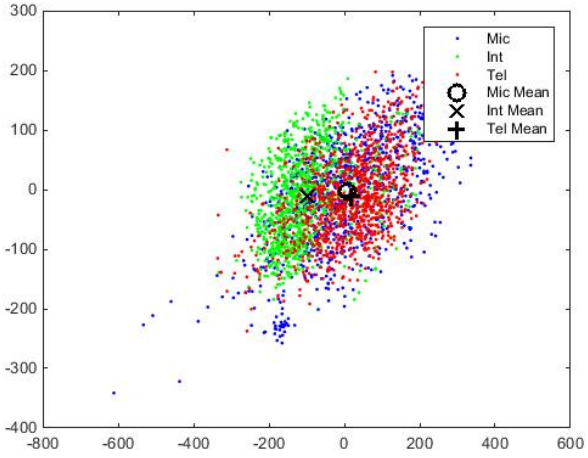


Fig. F.2: Distribution of i-vectors for three sources projected onto 2-dimensional space defined by the first two principal axes of PCA. Informative case.

microphone and interview data. Offsetting the prior mean[1] for each source in the i-vector extraction phase brings together the centers for all the sources.

[1]The reason for the slight difference between the mean of the interview data and that of the microphone data is due to the fact that only one prior is used.

4.2 Prior-compensated Total Variability Matrix Estimation

If we can further bring together the alignment of sources in the total variability space, then applying prior-compensated i-vector extraction in this new space ought to have the effect of applying a stronger source prior in the original space (note that the strength of the prior is dependent on the estimation method used). We propose to bring this alignment together in a single $\mathbf{T}$ matrix, where informative priors are utilized in the EM algorithm that is used in the training stage.

In the E-step, we fix $\mathbf{T}$, and for each source, find the posterior distribution of the latent variables that maximizes a pre-defined lower bound (a functional). In the M-step, we use the bound to find a more optimal value for $\mathbf{T}$. The EM algorithm guarantees that the log likelihood from $\mathbf{T}_i$ to $\mathbf{T}_{i+1}$ increases, where i represents the current training iteration. The proposed prior-compensated total variability matrix estimation is summarized as follows:

1. In the E-step, we first determine an informative prior $\mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma}_{\mathbf{p}})$ using minimum divergence estimation as described in Section 3.1. Using the informative prior, compute the posterior mean using (F.9) and covariance using (F.10).

2. In the M-step, accumulate the statistics across all sources, and use these to perform a single update for $\mathbf{T}$.

3. Repeat these EM steps until convergence.

Notice that step 2 above generates the prior-compensated i-vectors, as proposed in Section 3.2, together with their posterior covariance matrices. The source-related information is compensated for in the E-step. The resulting total variability matrix $\mathbf{T}$ could therefore be free from source-related influences. This is particularly useful when the amount of data is not balanced across data sources in a heterogeneous dataset. Alg. F.3 lists the details of prior-compensated EM.

5 Estimation Using Factor Analysis

In this section, we propose an estimation technique based on factor analysis. This can be particularly advantageous in the case when the microphone and interview data is more sparse than the telephone data, since covariance estimation using standard methods can in some cases lead to singular matrices. As we already know, given a set of Gaussian distributions, the Gaussian distribution that best represents these distributions can be obtained by minimizing the KL divergence between the two, and is given by the following objective function:

$$\Theta_{MD} = \sum_{i=1}^{I} \mathrm{E}\left[log \frac{\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{L}_i^{-1})}{\mathcal{N}(\mathbf{w}|\mu_{\mathbf{p}}, \mathbf{\Sigma_p}^{-1})} \right] \tag{F.14}$$

Algorithm F.3: Prior-compensated Expectation-Maximization

input : $\mathbf{T}, \mathbf{N}, \widetilde{\mathbf{F}}$
output: $\mathbf{T}$
begin
  // For each source, determine $\mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma_p})$
  $\mu_{\mathbf{p}} = \frac{1}{I} \sum_{i=1}^{I} \phi_i$
  $\mathbf{\Sigma_p} = \frac{1}{I} \sum_{i=1}^{I} (\phi_i - \mu_{\mathbf{p}})(\phi_i - \mu_{\mathbf{p}})^{\mathrm{T}} + \frac{1}{I} \sum_{i=1}^{I} \mathbf{L}_i^{-1}$
  // Reset accumulators
  $\mathbf{A} = \mathbf{0}$
  $\mathbf{C} = \mathbf{0}$
  // Expectation Step - Compute the posterior $p(\mathbf{w}|O) = \mathcal{N}(\phi,\ \mathbf{L}^{-1})$
  for $i = 1$ **to** I **do**
    $\mathbf{L}_i = \mathbf{T}^{\mathrm{T}} \mathbf{N}_i \mathbf{\Sigma}^{-1} \mathbf{T} + \mathbf{\Sigma_p}^{-1}$
    $\phi_i = \mathbf{L}^{-1} (\mathbf{T}^{\mathrm{T}} \mathbf{\Sigma}^{-1} \widetilde{\mathbf{F}}_{\mathbf{i}} + \mathbf{\Sigma_p}^{-1} \mu_{\mathbf{p}})$
    $E[\mathbf{w}_i \mathbf{w}_i^{\mathrm{T}}] = \phi_i {\phi_i}^{\mathrm{T}} + \mathbf{L}_i$
    // Acumulate Statistics
    $\mathbf{A} = \mathbf{A} + \mathbf{N}_i . E[\mathbf{w}_i \mathbf{w}_i^{\mathrm{T}}]$
    $\mathbf{C} = \mathbf{C} + \widetilde{\mathbf{F}}_i . \phi_i$
  end
  // Maximization Step - Solve the set of simultaneous equations: $\sum_i \mathbf{N}_i \mathbf{T} E[\mathbf{w}_i \mathbf{w}_i^{\mathrm{T}}] = \sum_i \mathbf{F}_i \phi_i$
  $\mathbf{T} = \mathbf{C} . \mathbf{A}^{-1}$ // Solves for $\mathbf{T}$
end

Note that this objective function is identical to that proposed in (12) of [25], but now with each posterior distribution characterized by its own separate covariance. We now wish to model these Gaussian distributions with a factor analyzer with mean $\mu_{\mathbf{p}}$ and covariance

$$\boldsymbol{\Sigma_{\mathbf{p}}}^{-1} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \mathbf{D} \tag{F.15}$$

Consider the following reformulation of the objective function:

$$\Theta_{MD} = \sum_{i=1}^{I} \mathrm{E}\left[log \frac{\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{L}_i^{-1})}{\mathcal{N}(\mathbf{w}|\mu_{\mathbf{p}}, \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \mathbf{D})} \right] \tag{F.16}$$

Since there is no closed-form solution for determining the parameters $\Theta = \{\mu_{\mathbf{p}}, \boldsymbol{\Phi}, \mathbf{D}\}$ for the factor analyzer, they can be learned using an EM algorithm, where the factor analyzer is expressed as a marginalization between the posterior distribution and a new hidden variable $\mathbf{h}_i$[2]. In the following M-step, the expectations are taken with respect to $\mathbf{w}$ and $\mathbf{h}_i$:

$$\Theta_{MD} = \sum_{i=1}^{I} \mathrm{E}_{\mathbf{w}}\left[log \frac{\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{L}_i^{-1})}{\mathrm{E}_{\mathbf{h}_i}[\mathcal{N}(\mathbf{w}|\mu_{\mathbf{p}} + \boldsymbol{\Phi}\mathbf{h}_i, \mathbf{D})]} \right] \tag{F.17}$$

In the E-step, we use standard methods to compute:

$$\begin{aligned} \mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i] =& (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\boldsymbol{\Phi} + \mathrm{I})^{-1} \cdot \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{m} - \mu_{\mathbf{p}}) \\ \mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i{\mathbf{h}_i}^{\mathrm{T}}] =& (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{D}^{-1}\boldsymbol{\Phi} + \mathrm{I})^{-1} + \mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}} \end{aligned} \tag{F.18}$$

Proposition 3: The update rules for the M step for Θ are given as:

$$\begin{aligned} \mu_{\mathbf{p}} =& \frac{1}{I}\sum_{i=1}^{I} \mathbf{m} \\ \boldsymbol{\Phi} =& \left[\sum_{i=1}^{I} (\mathbf{m} - \mu_{\mathbf{p}})\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\right] \cdot \left[\sum_{i=1}^{I} \mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i{\mathbf{h}_i}^{\mathrm{T}}]\right]^{-1} \\ \mathbf{D} =& \mathrm{diag}\left[\frac{1}{I}\sum_{i=1}^{I} (\mathbf{m} - \mu_{\mathbf{p}})(\mathbf{m} - \mu_{\mathbf{p}})^{\mathrm{T}} + \mathbf{L}_i^{-1} - \boldsymbol{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i](\mathbf{m} - \mu_{\mathbf{p}})^{\mathrm{T}}\right] \end{aligned} \tag{F.19}$$

The proof of the proposition is in the appendix.

Note the significance of combining factor analysis with minimum divergence estimation – it is simply a superposition of the individual approaches. The update formulas for $\mu_{\mathbf{p}}$ is identical to both the factor analysis and minimum divergence approach. The update formula for $\boldsymbol{\Phi}$ is identical to factor analysis, but not present

[2]Note that hidden variable $\mathbf{w}$ can be considered as observed, and simply corresponds to the already determined i-vector distribution as obtained in the previous sections.

with minimum divergence estimation. The update formula for $\mathbf{D}$ contains an empirical term common to both factor analysis and minimum divergence, a term identical to the posterior covariance in the minimum divergence approach and a residual term identical to the residual term in the factor analysis case.

6 Experiments

Experiments were carried out on the short2-short3 task of SRE'08 [10] and the core-core task of SRE'10 [11]. For all experiments, a gender-dependent setup was used, resulting in the training and use of separate male and female total variability matrices. The features used for training the 512-Gaussian UBMs were 57-dimensional MFCCs (including the first and second derivatives). The first-order statistics used for training each total variability matrix were centered and whitened [27].

In this work we compare four categories of systems, Category A, B, C and D. The results for all four categories are shown in Table F.1 and Table **??**. Category A is the baseline category and contains three baseline configurations: a telephone-only configuration, a pooled configuration and a cascade configuration. In the *telephone-only* configuration, a $\mathbf{T}$ matrix of dimension 600 was trained using only the telephone data. In the *pooled* configuration, a $\mathbf{T}$ matrix of dimension 600 was trained using pooled telephone and microphone data. In the *cascade* configuration, a $\mathbf{T}$ matrix of dimension 400 was trained using the telephone data, and a 200-dimensional supplementary $\mathbf{T}$ matrix was trained using microphone data [2]. The telephone data used to train these configurations was taken from SRE'04, 05 and 06. The microphone and interview data was taken from SRE'05, 06 and MIXER 5. For all three baseline configurations, the $\mathbf{T}$ matrices were trained using non-informative priors in the E-step. Following the training stage, 600-dimensional i-vectors were extracted using non-informative priors according to (F.5).

Category B comprises just a single configuration, where the already trained *pooled* $\mathbf{T}$ matrix was used as the fixed total variability matrix. Using minimum divergence estimation (Section 3.1), we trained one prior for the telephone subset and another prior for the microphone and interview subsets. The data used to train the telephone and microphone priors was identical to that outlined in Category A. We chose to use only one prior for both the microphone and interview cases since there was not enough interview data to reliably estimate a separate interview prior. 600-dimensional i-vectors were then extracted by performing re-centering of the first-order statistics using each prior's mean, followed by computation of the posteriors using each prior's informative covariance, as outlined in Section 4.

Category C contains four configurations, C1, C2, C3 and C4. Instead of using a pooled $\mathbf{T}$ matrix to extract i-vectors, a new 2-prior $\mathbf{T}$ matrix was trained using separately derived priors in the E-step, as outlined in section 4.2. Once again, these priors were extracted using minimum divergence estimation, with one prior for the

telephone subset and another for the microphone and interview subsets. The data used to train the telephone and microphone priors was identical to Category A and B. After training was complete, this new 2-prior $\mathbf{T}$ matrix was once again used to re-compute the priors needed for i-vector extraction, for all four configurations. For configuration C1, i-vectors were extracted using separate telephone and microphone priors. The priors were estimated using the minimum-divergence criterion. For C2, the telephone prior was estimated using the minimum divergence criterion and the microphone prior was estimated using factor analysis taking only the mean into account. Configuration C3 was almost identical to C2, except where now the microphone prior was estimated using factor analysis taking both the mean as well as the covariance into account. Finally, for C4, both the telephone and microphone priors were estimated using factor analysis taking both mean and covariance into account.

For the final category, Category D, the number of priors was increased to three, by using a separate prior for microphone and interview data. This category contains three configurations, D1, D2 and D3. In all three configurations, the 2-prior $\mathbf{T}$ matrix, used for Category C, was re-used here. For configuration D1, minimum divergence estimation was used to estimate each prior. For D2, factor analysis using only the mean of the posterior was used to estimate each prior. For D3, factor analysis using both the mean and covariance of the posterior were used to estimate the priors.

The following implementation details are common to all the above-mentioned experiments: The number of iterations used to train the $\mathbf{T}$ matrices was set to 20. For all priors trained using factor analysis, ten iterations were used for training the portrait matrix $\mathbf{\Phi}$ and $\mathbf{D}$, as outlined in Section 5. The number of factors used for the portrait matrix $\mathbf{\Phi}$ was set to 200. After extracting i-vectors, we used a similar experimental setup to [28] to carry out further processing and scoring. Firstly, LDA was used to reduce the dimension of the i-vectors from 600 to 400. After length normalization [3], PLDA was used to model the channel variability [29]. For the PLDA model, a 200-dimensional telephone matrix was trained using telephone data, and a 50-dimensional microphone matrix was trained, using microphone and interview data. The matrices were trained in a decoupled manner.

7 Discussion

We present results for the all four categories, for both male and female trials. For all results, we show both Equal Error Rate (EER) as well as the Detection Cost Function (DCF) (DCF-08 for SRE'08 results and DCF-10 for SRE'10 results). The SRE'08 results are shown in Table F.1, and the SRE'10 results are shown in Table F.2. For each condition, results for both female and male trials are shown.

Table F.1: SRE'08 performance comparison for the eight common conditions (CC) in the short2-short3 core task. Left: FEMALE Trials, Right: MALE Trials. Each entry consists of EER (top) and DCF-08 (bottom). Entries in **bold** either match or beat the 3 baselines in Category A. Starred entries (*) in rows C1-C4 shown for cases that outperform B1. Plus entries (+) shown where D3 outperforms D1 and D2. NI is Non-informative prior (only 1).

			CC1: int-int		CC2: int-int		CC3: int-int		CC4: int-tel		CC5: tel-mic		CC6: tel-tel		CC7: tel-tel		CC8: tel-tel	
Configuration	T	I-vector	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
A1	Telephone-only	NI	3.51 0.16	2.84 0.13	1.50 0.05	0.32 0.00	3.61 0.17	2.97 0.13	5.69 0.28	4.07 0.19	6.65 0.27	4.17 0.19	5.85 0.31	4.67 0.25	2.73 0.14	2.32 0.11	3.24 0.15	1.43 0.06
A2	Pooled	NI	3.22 0.16	2.54 0.13	1.28 0.06	0.33 0.01	3.29 0.16	2.64 0.13	4.65 0.25	3.89 0.19	5.62 0.26	3.05 0.14	5.86 0.29	4.15 0.25	2.84 0.14	1.60 0.11	3.32 0.14	1.04 0.07
A3	Cascade	NI	3.17 0.16	3.01 0.14	1.25 0.05	0.41 0.02	3.27 0.16	3.22 0.15	5.38 0.26	4.27 0.20	6.10 0.26	4.12 0.16	5.86 0.30	4.06 0.23	2.98 0.13	1.66 0.10	3.81 0.15	1.32 0.07
B1	Pooled	2-priors	**2.34** **0.12**	**1.95** **0.09**	1.32 0.06	**0.32** **0.00**	**2.39** **0.12**	**2.04** **0.10**	**4.32** **0.23**	3.91 0.20	**5.37** **0.26**	3.21 0.17	**5.79** **0.29**	**3.84** 0.24	2.87 **0.13**	**1.39** 0.11	3.27 **0.12**	**0.90** 0.07
C1	2-priors	2-priors	**2.33*** **0.12**	**1.96** **0.10**	**1.20*** **0.05***	**0.24*** **0.00**	**2.37*** **0.12**	**2.04** **0.10**	**4.24*** **0.22***	3.90* **0.19***	**5.07*** **0.23***	3.09* 0.15*	**5.72*** **0.28***	**3.67*** 0.24	2.80* **0.13**	**1.42** 0.11	3.29 **0.13**	**0.87*** 0.07
C2	2-priors	2-priors	**2.26*** **0.12**	**1.99** **0.10**	1.58 0.06	**0.18*** 0.01	**2.28*** **0.13**	**2.09** **0.11**	**4.20*** **0.21***	3.94 0.20	**4.83*** **0.23***	**2.97*** 0.16*	**5.68*** **0.29**	**3.82*** 0.24	2.74* **0.13**	**1.56** 0.11	3.42 **0.14**	**0.76*** 0.07
C3	2-priors	2-priors	**2.25*** **0.12**	**1.99** **0.10**	1.56* 0.06	**0.27*** **0.00**	**2.29*** **0.12**	**2.09** **0.10**	**4.24*** **0.21***	**3.86*** 0.20	**5.09*** **0.23***	3.08* 0.15*	**5.68*** **0.29**	**3.94** 0.24	2.88 **0.13**	**1.47** 0.10	3.41 **0.13**	**0.81*** **0.06**
C4	2-priors	2-priors	**2.40** **0.13**	**2.01** **0.10**	1.60 0.06	**0.27*** **0.00**	**2.44** **0.13**	**2.10** **0.10**	**4.05*** **0.21***	3.98 0.20	**5.01*** **0.24***	3.06* 0.15*	**5.70*** **0.28***	**3.79*** 0.24	2.84* **0.13**	**1.35*** 0.11	3.34 **0.14**	**0.78*** 0.07
D1	2-priors	3-priors	4.57 0.20	3.18 0.15	2.50 0.06	0.41 0.01	4.69 0.21	3.33 0.16	7.33 0.37	4.59 0.24	5.15 0.24	3.24 0.14	5.67 0.29	3.72 0.23	2.83 0.14	1.43 0.10	3.26 0.14	0.66 0.07
D2	2-priors	3-priors	3.40 0.16	2.67 0.13	1.99 0.04	0.41 0.02	3.48 0.17	2.78 0.14	6.28 0.34	4.14 0.21	5.06 0.23	3.26 0.14	5.69 0.29	3.80 0.24	2.80 0.13	1.50 0.11	3.44 0.14	0.85 0.07
D3	2-priors	3-priors	**3.00+** **0.15+**	**2.52+** **0.13**	1.73+ **0.04**	**0.27+** 0.03	**3.07+** **0.16+**	**2.64+** 0.14	5.77+ 0.31+	4.18 0.20+	**5.05+** **0.24**	3.27 0.15	**5.71** **0.28+**	**3.95** 0.25	2.83 **0.13**	**1.52** 0.11	3.46 **0.13+**	**0.73+** 0.07

Table F.2: SRE'10 performance comparison for the nine common conditions (CC) for the core-core core task. Left: FEMALE Trials, Right: MALE Trials. Each entry consists of EER (top) and DCF-10 (bottom). Entries in **bold** either match or beat the 3 baselines in Category A. Starred entries (*) in rows C1-C4 shown for cases that outperform B1. Plus entries (+) shown where D3 outperforms D1 and D2. NI is Non-informative prior (only 1).

			CC1: int-int-same-mic		CC2: int-int-diff-mic		CC3: int-tel		CC4: int-mic		CC5: nve-nve-diff-tel		CC6: nve-hve-diff-tel		CC7: nve-hve-mic		CC8: nve-lve-diff-tel		CC9: nve-lve-mic	
Configuration	T	I-vector	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
A1	Telephone-only	NI	3.06 0.36	2.02 0.22	5.65 0.67	3.45 0.47	4.21 0.57	3.55 0.50	3.96 0.54	2.72 0.36	3.59 0.40	3.47 0.46	8.09 0.79	4.64 0.70	8.49 0.69	4.91 0.46	2.01 0.43	1.14 0.23	2.46 0.27	1.54 0.22
A2	Pooled	NI	3.16 0.34	2.22 0.20	5.13 0.60	3.14 0.40	3.34 0.55	2.82 0.40	3.78 0.50	2.54 0.36	3.00 0.37	2.60 0.45	7.13 0.79	4.01 0.67	7.98 0.70	4.95 0.45	1.66 0.29	1.54 0.23	2.55 0.23	1.34 0.24
A3	Cascade	NI	3.12 0.38	2.29 0.24	5.60 0.61	3.29 0.43	4.01 0.52	2.62 0.49	4.04 0.51	2.87 0.33	3.41 0.34	3.13 0.36	7.10 0.82	4.33 0.74	8.19 0.74	5.25 0.45	1.83 0.29	1.68 0.26	3.08 0.30	1.61 0.21
B1	Pooled	2-priors	**2.43** 0.35	**1.67** 0.24	**4.44** 0.62	**2.25** 0.43	3.87 0.62	3.19 0.61	**3.33** **0.49**	**2.22** 0.36	**3.00** 0.42	2.89 0.39	7.11 **0.77**	4.13 0.69	**7.49** **0.59**	**4.16** 0.51	**1.59** 0.34	1.56 0.26	2.48 0.31	**1.15** 0.24
C1	2-priors	2-priors	**2.56** **0.24***	**1.48*** 0.24	**4.54** **0.45***	**2.20*** 0.45	3.86* 0.60*	2.87* 0.59*	**3.32*** **0.38***	**2.12*** 0.38	3.13 0.40*	2.82* 0.40	**7.19** **0.67***	4.39 **0.67***	**7.69** **0.48***	**4.67** **0.48***	**1.65** **0.23***	1.51* **0.23***	2.65 0.25*	**0.85*** 0.26
C2	2-priors	2-priors	**2.67** 0.37	**1.69** 0.26	**4.62** **0.57***	**2.54** **0.39***	3.75* 0.60*	3.07* 0.58*	**3.28*** **0.49**	**2.24** 0.36	3.26 0.41*	2.76* 0.38*	**7.18** 0.80	4.18 **0.66***	**7.57** **0.59**	**4.61** **0.42***	**1.61** 0.36	1.47* **0.18***	2.68 0.27*	**1.16** 0.28
C3	2-priors	2-priors	**2.63** 0.36	**1.81** 0.24	**4.62** **0.58***	**2.58** 0.42*	3.83* 0.63	3.02* 0.52*	**3.39** **0.49**	**2.27** **0.33***	3.26 0.40*	2.78* 0.40	**7.08*** **0.75***	4.10* **0.64***	**7.37*** **0.57***	**4.72** **0.42***	**1.67** 0.38	1.34* **0.18***	2.70 0.27*	**1.34** **0.21***
C4	2-priors	2-priors	**2.61** 0.36	**1.79** 0.24	**4.60** **0.56***	**2.53** 0.42*	3.79* 0.60*	3.09* 0.57*	**3.29*** 0.53	**2.30** 0.34*	3.12 0.42	2.90 0.39	**7.22** 0.80	4.01* 0.69	**7.81** **0.58***	**4.48** **0.44***	**1.62** 0.38	1.62 **0.18***	2.64 0.27*	**1.27** 0.26
D1	2-priors	3-priors	4.48 0.48	2.40 0.34	8.45 0.78	3.41 0.53	6.87 0.74	4.12 0.66	4.98 0.66	2.87 0.38	3.03 0.41	2.88 0.41	7.32 0.74	4.09 **0.63**	**7.71** **0.67**	**4.15** **0.42**	1.67 0.39	1.64 **0.19**	2.61 0.29	**0.89** 0.24
D2	2-priors	3-priors	3.82 0.38	2.05 0.27	6.58 0.74	**3.01** 0.49	6.23 0.75	4.14 0.69	4.74 0.66	3.00 0.43	3.23 0.41	2.62 0.41	**6.98** 0.80	4.19 0.70	8.12 **0.66**	**4.55** **0.39**	**1.63** 0.38	1.65 **0.22**	2.71 0.27	1.37 0.25
D3	2-priors	3-priors	3.60+ 0.37+	2.07 0.28	6.11+ 0.69+	**2.95+** 0.48+	5.69+ 0.70+	4.02+ 0.61+	4.56+ 0.63+	2.76+ 0.40	3.20 0.39+	2.73 0.40+	**6.69+** **0.79**	**3.95+** 0.69	**7.90** **0.64+**	**4.46** **0.39**	**1.63** 0.37+	1.64 **0.18+**	2.54+ 0.26+	**1.29** 0.26

7.1 2-prior Compensated i-vector Extraction from Pooled Total Variability Matrix

We begin by discussing Category B, the first proposed configuration. Looking at the SRE'08 results with respect to EER for B1, we note a strong improvement in common conditions 1 and 3, corresponding to the *int-int* condition. We could not beat the baseline for CC 2, which we believe is due to the smaller number of trials. For the mixed trials, i.e. CCs 4 and 5, source-specific informative priors showed improved robustness against both the telephone-only and cascade cases. For the pooled case however, the results were a lot closer and we did not beat this baseline in all cases. We observed the best results were found for the female trials, where all baselines were improved upon. Interestingly, our approach improved on several of the *tel-tel* only conditions, more notably in the male case. From these results, it appears that source-specific informative priors offer the greatest strength in enhancing performance trials where the sources of the trial and target match. For the B1 results, for both the female and male tasks, we were able to improve on six out of eight conditions.

We also examine the SRE'10 results, again with respect to their EER. For the single source interview and mic common conditions, as given by CCs 1, 2, 7 and 9, we were able to beat all baselines in three out of four CCs in the female case and all CCs in the male case. For telephone-only trials, given by CCs 5, 6 and 8, in only one instance could all baselines be beaten. We believe the reason for the slightly worse results for SRE'10 is the decreased similarity of the data used to train the $\mathbf{T}$-matrices and subspace PLDA models to that of SRE'08. For the cross-channel conditions, we noted better performance for the *int-mic* cross channel than for *int-tel*, strengthening our belief that best performance is gained where source and target trials more closely match one another. For the female sub-task, we were able to beat 6 out 9 conditions and for the male sub-task, performance was improved for 5 out of 9 conditions.

7.2 2-prior Compensated i-vector Extraction using 2-prior Compensated Total Variability Matrix

In section 4.2, we saw how it is possible to incorporate informative priors into the construction of the total variability matrix in addition to the i-vector extraction stage. When using such a configuration with just two informative priors, using a variety of estimation techniques, we obtain the results shown in Category C. Starred entries (*) indicate trials where the performance was better than using a pooled $\mathbf{T}$ matrix (B1).

According to the EER shown for the SRE'08 results, for C1, six out of eight conditions beat the baselines (the same as the number beaten in the pooled $\mathbf{T}$ approach) - however the conditions where this occurred were slightly different. Generally, we noted an improvement over B1 in seven out of eight common condi-

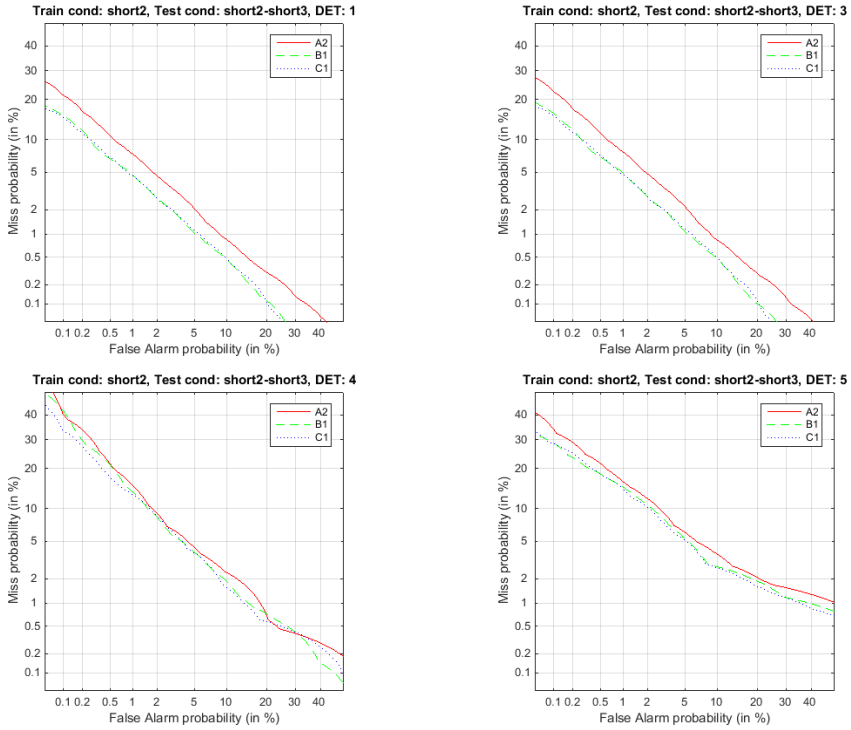


Fig. F.3: SRE'08 Detection Error Tradeoff (DET) curves for the four common conditions (CC) 1, 3, 4 and 5 for the short2-short3 core task.

tions for female trials, and five out of eight common conditions for male trials. In particular, a much better performance is seen for CC 2, which for B1, was unable to beat the baseline. For the remaining configurations C2, C3 and C4, which used factor analysis with varying degrees, we did not see a large deviation in the results compared to C1. This we believe is due to the fact that combining the microphone and interview data into a single prior reduces the sparseness of such a prior, and renders factor analysis for estimating such a prior less effective compared to minimum divergence estimation. The SRE-10 results paint a similar picture, where for C1, six out of nine for female trials and five out of nine for male trials were able to beat the baselines. In only two cases was there an performance improvement over C1 for female trials and six for male trials. Interestingly, this was significantly higher when considering the DCF-10, where an improvement was seen in C1 over B1 for every female trial and four male trials. Once again, we did not see a large enough deviation in C2, C3 and C4 to warrant using factor analysis for estimating priors over minimum divergence.

Fig. F.3 shows superimposed Detection Error Tradeoff (DET) [30] curves for the three conditions A2, B1 and C2. For the *int-int* CCs 1 and 3 we notice a fairly

constant improvement, with a substantial improvement from A2 to B1, and almost similar performance for B1 and C1. For the mixed conditions, i.e. CC 4 and 5 there is still an improvement of B1 over A2, but this is less marked, and in the case of CC 4, we noticed for a small range of threshold values, that A2 performed the best. For the CCs B1 and C1, we see that in general the performance of one over the other is very much dependent on the threshold setting that is chosen, but from visual inspection, C1 appears to perform B1 for the mixed channel conditions.

In general, it appears from the above results that incorporating prior modeling into the estimation of the $\mathbf{T}$ matrix leads to a better initial estimation for $\mathbf{T}$, and we believe this has the effect of compensating for source variation even before considering using informative priors for i-vector extraction. We believe that using the same $\mathbf{T}$ matrix for i-vector estimation is equivalent to increasing the strength of the prior and results in more effective overall source compensation.

Finally, we discuss the results for Category D, that relates to re-using the $\mathbf{T}$ matrix from Category B and C, but where three priors are used for i-vector extraction. Once again we present the SRE-08 EER results first. Firstly, we note from D1, where minimum divergence was used to estimate each prior, that the performance is generally a lot worse than all other cases - most likely due to the increased sparsity of the data across three priors instead of two. When factor analysis is used, by taking only the mean into account, to estimate the priors instead, we notice an improvement in six out of eight (female) and three out of eight (male). Two things are apparent here: firstly, where there is improvement, it is substantial, and secondly, improvement is mostly seen in the microphone / interview trials and cross trials. This second observation is expected, since there is no change in the sparsity of the telephone data (the subset is the same in the 2-prior and 3-prior case), and hence we do not expect to see an improvement with factor analysis (as was also observed under category C). When using factor analysis, but taking also the covariance into account as well as the mean, as shown in D3, we notice improvements over D2 in five out of eight (female) and four out of eight (male). Once again, these were mostly observed for the microphone / interview trials and mixed channel trials. Interestingly, four female and six male conditions were able to beat the baseline in D3, as opposed to none for both D1 and D2, showing the effectiveness of factoring in the covariance as well into the factor analysis estimation technique. For the SRE-10 data, improvements were seen in D2 over D1 in six out of nine conditions (female) and three out of nine conditions (male). Similarly, from D3, further improvements were seen over D2 in six out of nine (female) and four out of nine (male) conditions. Here, three female and male trials were able to beat all baselines. Applying factor analysis appears to work well in cases where the data is sparse, but not so well where the data is abundant.

8 Conclusion

In this paper, we proposed a novel method of incorporating informative priors into the posterior computation used in total variability modeling. The purpose of doing this is to better describe the source variation from a heterogeneous dataset. We propose to estimate informative priors using minimum divergence estimation and using these priors to compute the posteriors at both the i-vector extraction stage as well as the E-step in training a $\mathbf{T}$ matrix. We showed that both strategies have a positive influence on speaker recognition performance. Conducting i-vector extraction using two-source priors using an already trained matrix of pooled data led to performance gains in six out of eight common conditions for the short2-short3 task in SRE'08 for both genders, and six out of nine female and five out of nine male common-conditions for the core-core task in SRE'10. Using the source priors in the E-step to train a new $\mathbf{T}$ matrix, and once again carrying out 2-prior i-vector extraction led to performance gains in six out of eight common-conditions for the short2-short3 core task in SRE'08 for both genders, and six out of nine female and five out of nine male common conditions for the core-core core task in SRE'10.

We also investigated the use of both minimum divergence and factor analysis for the 3-prior case. In the case of minimum divergence, we found the performance was rather poor, which we believe to be due to the sparsity of the data. However, factor analysis for the three-prior case showed very promising results, as opposed to minimum divergence. Using only the mean showed a dramatic performance improvement – for SRE'08 this improvement was seen in six female and eight male common conditions, and for SRE'10 the improvement was in nine female and three male conditions. Considering both the mean and covariance showed an additional improvement of five female and four male conditions for SRE'08 and six female and four male. However, due to the sparsity of the data for the two-prior case, only some conditions were able to beat the reference baselines.

A Proofs of the Propositions

A.1 Proposition 1

Proof. Assume that we have the parameter set $\{\mathbf{T}, \mathbf{\Sigma}\}$, a set of hidden variables $\mathbf{w}$ and a sequence of observations O. From Lemma 1 in [21] we know that the log likelihood of O given $\mathbf{w}$ for the parameters $\{\mathbf{T}, \mathbf{\Sigma}\}$ can be expressed as the sum of two terms:

$$\log P_{\mathbf{T},\mathbf{\Sigma}}(O|\mathbf{w}) = \mathrm{G}_{\mathbf{T}} + \mathrm{H}_{\mathbf{T},\mathbf{\Sigma}} \tag{F.20}$$

where $G_{\mathbf{T}}$ is defined by (3) in [21], and $\mathrm{H}_{\mathbf{T},\mathbf{\Sigma}}$ is defined as

$$H_{\mathbf{T},\mathbf{\Sigma}} = \mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\widetilde{\mathbf{F}} - \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{N}\mathbf{\Sigma}^{-1}\mathbf{T}\mathbf{w} \tag{F.21}$$

Since we are primarily interested in estimating $\mathbf{T}$ and not $\mathbf{\Sigma}$, and since $\mathrm{G}_{\mathbf{T}}$ does not depend on $\mathbf{T}$, this term is not considered further [21]. Given the mean $\mu_{\mathbf{p}}$ and covariance ${\mathbf{\Sigma}_{\mathbf{p}}}^{-1}$ for the prior, we can express the prior as:

$$P(\mathbf{w}) \propto exp(-\frac{1}{2}(\mathbf{w}-\mu_{\mathbf{p}})^{\mathrm{T}}{\mathbf{\Sigma}_{\mathbf{p}}}^{-1}(\mathbf{w}-\mu_{\mathbf{p}})) \tag{F.22}$$

The posterior distribution for $\mathbf{w}$ given $\mathrm{H}_{\mathbf{T},\mathbf{\Sigma}}$ is formulated as

$$\begin{aligned} P(\mathbf{w}|O) = exp(&\mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{F} - \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{N}\mathbf{\Sigma}^{-1}\mathbf{T}\mathbf{w} - \\ &\frac{1}{2}(\mathbf{w}-\mu_{\mathbf{p}})^{\mathrm{T}}{\mathbf{\Sigma}_{\mathbf{p}}}^{-1}(\mathbf{w}-\mu_{\mathbf{p}})) \\ \propto exp(&-\frac{1}{2}(\mathbf{w}-\mu)^{\mathrm{T}}\mathbf{P}(\mathbf{w}-\mu)) \end{aligned} \tag{F.23}$$

with $\mathbf{P}$ and μ in the stated form as above. □

A.2 Proposition 2

Proof. We first derive the probability distribution of $p(\mathbf{m})$ where $\mathbf{m} = \mathbf{m_0} + \mathbf{T}\mathbf{w}$ and $\mathbf{w} \sim \mathcal{N}(\mu_{\mathbf{p}}, \mathbf{\Sigma_p})$. The mean is computed as:

$$\begin{aligned} \mathrm{E}[\mathbf{m}] &= \mathrm{E}[\mathbf{m_0} + \mathbf{T}\mathbf{w}] \\ &= \mathrm{E}[\mathbf{m_0}] + \mathrm{E}[\mathbf{T}\mathbf{w}] \\ &= \mathbf{m_0} + \mathbf{T}\mu_{\mathbf{p}} \end{aligned} \tag{F.24}$$

and covariance as:

$$\begin{aligned} \mathrm{E}[(\mathbf{m}-\mathrm{E}[\mathbf{m}])^2] =& \mathrm{E}[(\mathbf{m_0} + \mathbf{T}\mathbf{w} - \mathbf{m_0} - \mathbf{T}\mu_{\mathbf{p}}). \\ & (\mathbf{m_0} + \mathbf{T}\mathbf{w} - \mathbf{m_0} - \mathbf{T}\mu_{\mathbf{p}})^{\mathrm{T}}] \\ =& \mathbf{T}\mathrm{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]\mathbf{T}^{\mathrm{T}} - \mathbf{T}\mu_{\mathbf{p}}{\mu_{\mathbf{p}}}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}} \end{aligned} \tag{F.25}$$

We continue the derivation to find an expression for $E[\mathbf{w}\mathbf{w}^{\mathrm{T}}]$ for the informative case. In this case we know that the covariance of the prior distribution for $p(\mathbf{w})$ is simply $\mathbf{\Sigma_p}$, and so we have:

$$\mathbf{\Sigma_p} = \mathrm{E}[(\mathbf{w}-\mathrm{E}[\mathbf{w}])^2] = \mathrm{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] - \mu_{\mathbf{p}}{\mu_{\mathbf{p}}}^{\mathrm{T}} \tag{F.26}$$

Substituting the results of (F.26) into (F.25) gives:

$$\begin{aligned} \mathrm{E}[(\mathbf{m}-\mathrm{E}[\mathbf{m}])^2] =& \mathbf{T}(\mathbf{\Sigma_p} + \mu_{\mathbf{p}}{\mu_{\mathbf{p}}}^{\mathrm{T}})\mathbf{T}^{\mathrm{T}} - \mathbf{T}\mu_{\mathbf{p}}{\mu_{\mathbf{p}}}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}} \\ =& \mathbf{T}\mathbf{\Sigma_p}\mathbf{T}^{\mathrm{T}} \end{aligned} \tag{F.27}$$

Note that for the case of the non-informative prior, the mean and covariance reduce to $\mathbf{m_0}$ and $\mathbf{T}\mathbf{T}^{\mathrm{T}}$, respectively. In the same vein, we compute the mean for the marginalization modeled by $\mathbf{m} = \mathbf{m_0} + \mathbf{T}\mathbf{w} + \mathbf{T}\mu_{\mathbf{p}}$ and $\mathbf{w} \sim \mathcal{N}(0, \mathbf{\Sigma_p})$. We find the mean to be

$$\begin{aligned} \mathrm{E}[\mathbf{m}] &= \mathrm{E}[\mathbf{m_0} + \mathbf{T}\mathbf{w} + \mathbf{T}\mu_{\mathbf{p}}] \\ &= \mathbf{m_0} + \mathbf{T}\mu_{\mathbf{p}} \end{aligned} \tag{F.28}$$

which is identical to the formally derived mean in (F.24). The covariance is computed as:

$$\begin{aligned} \mathrm{E}[(\mathbf{m} - \mathrm{E}[\mathbf{m}])^2] =& \mathrm{E}[(\mathbf{m_0} + \mathbf{T}\mathbf{w} + \mathbf{T}\mu_{\mathbf{p}} - \mathbf{m_0} - \mathbf{T}\mu_{\mathbf{p}}). \\ & (\mathbf{m_0} + \mathbf{T}\mathbf{w} + \mathbf{T}\mu_{\mathbf{p}} - \mathbf{m_0} - \mathbf{T}\mu_{\mathbf{p}})^{\mathrm{T}}] \\ =& \mathbf{T}\mathrm{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]\mathbf{T}^{\mathrm{T}} \end{aligned} \tag{F.29}$$

To proceed we relate $\mathbf{\Sigma_p}$ and $\mathrm{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]$ to one another:

$$\mathbf{\Sigma_p} = \mathrm{E}[(\mathbf{w} - \mathrm{E}[\mathbf{w}])^2] = \mathrm{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] \tag{F.30}$$

Substituting the result of (F.30) into (F.29) gives:

$$\mathrm{E}[(\mathbf{m} - \mathrm{E}[\mathbf{m}])^2] = \mathbf{T}\mathbf{\Sigma_p}\mathbf{T}^{\mathrm{T}} \tag{F.31}$$

□

which is identical to the formally derived covariance in (F.27). These will contribute equally to the marginalization $\Pi(c)$ given in (F.11). This concludes the proof.

A.3 Proposition 3

Proof. The objective function for the factor analyzer as given by (F.17) can be broken down as follows:

$$\begin{aligned} \Theta_{MD} = \sum_{i=1}^{I} \mathrm{E}_{\mathbf{w}} \Bigg[& -\frac{1}{2}[\log|\mathbf{L}_i^{-1}| + (\mathbf{w} - \mathbf{m})^{\mathrm{T}}\mathbf{L}_i(\mathbf{w} - \mathbf{m})] - \\ & \mathrm{E}_{\mathbf{h}_i}\Big[\frac{1}{2}\log|\mathbf{\Sigma}| + (\mathbf{w} - \mu_{\mathbf{p}} - \mathbf{\Phi}\mathbf{h}_i)^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{w} - \mu_{\mathbf{p}} - \mathbf{\Phi}\mathbf{h}_i)\Big]\Bigg] \end{aligned} \tag{F.32}$$

Considering momentarily only the last part of (F.32) relating to the denominator in (F.17), we have:

$$\begin{aligned}
&\mathrm{E}_{\mathbf{h}_i}[\log|\mathbf{\Sigma}| + (\mathbf{w}-\mu_\mathbf{p}-\mathbf{\Phi}\mathbf{h}_i)^\mathrm{T}\mathbf{D}^{-1}(\mathbf{w}-\mu_\mathbf{p}-\mathbf{\Phi}\mathbf{h}_i)] \\
&= \log|\mathbf{\Sigma}| + \mathbf{w}^\mathrm{T}\mathbf{D}^{-1}\mathbf{w} + \mathbf{y}^\mathrm{T}\mathbf{D}^{-1}\mu_\mathbf{p} + \\
&\quad \mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^\mathrm{T}\mathbf{\Phi}^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i] - 2\mathbf{w}^\mathrm{T}\mathbf{D}^{-1}\mu_\mathbf{p} - \\
&\quad 2\mathbf{w}^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i] + 2\mathbf{y}^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]
\end{aligned} \tag{F.33}$$

Plugging (F.33) back into (F.32) and expanding the first part, we have:

$$\begin{aligned}
\Theta_{MD} = &\sum_{i=1}^{I} \mathrm{E}_\mathbf{w}\Bigg[-\frac{1}{2}\Big[\log|\mathbf{L}_i^{-1}| + \mathbf{w}^\mathrm{T}\mathbf{L}_i\mathbf{w} + \mathbf{m_r}^\mathrm{T}\mathbf{L}_i\mathbf{m} - \\
&2\mathbf{w}^\mathrm{T}\mathbf{L}_i\mathbf{m} - \log|\mathbf{\Sigma}| - \mathbf{w}^\mathrm{T}\mathbf{D}^{-1}\mathbf{w} - \mathbf{y}^\mathrm{T}\mathbf{D}^{-1}\mu_\mathbf{p} - \\
&\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^\mathrm{T}\mathbf{\Phi}^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i] + 2\mathbf{w}^\mathrm{T}\mathbf{D}^{-1}\mu_\mathbf{p} + \\
&2\mathbf{w}^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i] - 2\mathbf{y}^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]\Big]\Bigg] \\
= &\sum_{i=1}^{R} -\frac{1}{2}(\log|\mathbf{L}_i^{-1}| + \mathrm{E}_\mathbf{w}[\mathbf{w}]^\mathrm{T}\mathbf{L}_i\mathrm{E}_\mathbf{w}[\mathbf{w}] + \mathbf{m_r}^\mathrm{T}\mathbf{L}_i\mathbf{m} - \\
&2\mathrm{E}_\mathbf{w}[\mathbf{w}]^\mathrm{T}\mathbf{L}_i\mathbf{m} - \log|\mathbf{\Sigma}| - \mathrm{E}_\mathbf{w}[\mathbf{w}]^\mathrm{T}\mathbf{D}^{-1}\mathrm{E}_\mathbf{w}[\mathbf{w}] - \\
&\mathbf{y}^\mathrm{T}\mathbf{D}^{-1}\mu_\mathbf{p} - \mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^\mathrm{T}\mathbf{\Phi}^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i] + \\
&2\mathrm{E}_\mathbf{w}[\mathbf{w}]^\mathrm{T}\mathbf{D}^{-1}\mu_\mathbf{p} + 2\mathrm{E}_\mathbf{w}[\mathbf{w}]^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i] - \\
&2\mathbf{y}^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i])
\end{aligned} \tag{F.34}$$

We are now in a position to compute the update rules. Taking the derivative of (F.34) with respect to $\mu_\mathbf{p}$ gives the following:

$$\frac{\partial}{\partial\mu_\mathbf{p}}\Theta_{MD} = \sum_{i=1}^{I} -\frac{1}{2}(-2\mathbf{D}^{-1}\mu_\mathbf{p} + 2\mathbf{D}^{-1}\mathrm{E}_\mathbf{w}[\mathbf{w}] - 2\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]) \tag{F.35}$$

$$\therefore \sum_{i=1}^{R} \mu_\mathbf{p} - \mathrm{E}_\mathbf{w}[\mathbf{w}] + \mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i] = 0 \tag{F.36}$$

$$\therefore R\mu_\mathbf{p} = \sum_{i=1}^{R} \mathrm{E}_\mathbf{w}[\mathbf{w}] - \mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i] \tag{F.37}$$

Realizing that $\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i] = (\mathbf{\Phi}^\mathrm{T}\mathbf{D}^{-1}\mathbf{\Phi} + \mathrm{I})^{-1}\mathbf{\Phi}^\mathrm{T}\mathbf{D}^{-1}(\mathrm{E}_\mathbf{w}[\mathbf{w}] - \mu_\mathbf{p})$, and the $\mathrm{E}_\mathbf{w}[\mathbf{w}] = \mathbf{m}$, the above expression simplifies to

$$\mu_\mathbf{p} = \frac{1}{I}\sum_{i=1}^{I} \mathbf{m} \tag{F.38}$$

which is the expected expression for the mean value of $\mu_{\mathbf{p}}$.

We now proceed to determine the update rule for $\mathbf{\Phi}$ by evaluating the derivative (F.34) with respect to $\mathbf{\Phi}$:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{\Phi}}\Theta_{MD} = \sum_{i=1}^{I} -\frac{1}{2}(-2\mathbf{D}^{-1}\mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i\mathbf{h}_i{}^{\mathrm{T}}]+ \\ 2\mathbf{D}^{-1}\mathrm{E}_{\mathbf{w}}[\mathbf{w}]\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}} - 2\mathbf{D}^{-1}\mu_{\mathbf{p}}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}})\end{aligned}$$

$$\therefore \sum_{i=1}^{I} \mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i\mathbf{h}_i{}^{\mathrm{T}}] - \mathrm{E}_{\mathbf{w}}[\mathbf{w}]\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}} + \mu_{\mathbf{p}}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}} = 0 \tag{F.39}$$

$$\therefore \sum_{i=1}^{I} \mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i\mathbf{h}_i{}^{\mathrm{T}}] = \sum_{i=1}^{I}(\mathrm{E}_{\mathbf{w}}[\mathbf{w}] - \mu_{\mathbf{p}})\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}} \tag{F.40}$$

Solving for $\mathbf{\Phi}$ gives:

$$\begin{aligned}\therefore \mathbf{\Phi} &= \left[\sum_{i=1}^{I}(\mathrm{E}_{\mathbf{w}}[\mathbf{w}] - \mu_{\mathbf{p}})\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\right].\left[\sum_{i=1}^{I}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i\mathbf{h}_i{}^{\mathrm{T}}]\right]^{-1} \\ &= \left[\sum_{i=1}^{I}(\mathbf{m} - \mu_{\mathbf{p}})\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\right].\left[\sum_{i=1}^{I}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i\mathbf{h}_i{}^{\mathrm{T}}]\right]^{-1}\end{aligned} \tag{F.41}$$

Finally we determine the update rule for $\mathbf{D}$ by evaluating the derivative of (F.34) with respect to $\mathbf{D}$:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{D}}\Theta_{MD} = \sum_{i=1}^{I} -\frac{1}{2}\Big[-\mathbf{D}^{-1} - \mathbf{D}^{-1}(-\mathrm{E}_{\mathbf{w}}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] - \mu_{\mathbf{p}}\mathbf{y}^{\mathrm{T}}- \\ \mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i\mathbf{h}_i{}^{\mathrm{T}}]\mathbf{\Phi}^{\mathrm{T}} + 2\mathrm{E}_{\mathbf{w}}[\mathbf{w}]\mathbf{y}^{\mathrm{T}}+ \\ 2\mathrm{E}_{\mathbf{w}}[\mathbf{w}]\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}} - 2\mu_{\mathbf{p}}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}})\mathbf{D}^{-1}\Big]\end{aligned} \tag{F.42}$$

$$\begin{aligned}\therefore \sum_{i=1}^{I} -\frac{1}{2}(-\mathbf{D}^{-1} - \mathbf{D}^{-1}(-\mathrm{E}_{\mathbf{w}}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] - \mu_{\mathbf{p}}\mathbf{y}^{\mathrm{T}}- \\ \mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i\mathbf{h}_i{}^{\mathrm{T}}]\mathbf{\Phi}^{\mathrm{T}} + 2\mathrm{E}_{\mathbf{w}}[\mathbf{w}]\mathbf{y}^{\mathrm{T}} + 2\mathrm{E}_{\mathbf{w}}[\mathbf{w}]\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}- \\ 2\mu_{\mathbf{p}}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}})\mathbf{D}^{-1}) = 0\end{aligned} \tag{F.43}$$

Solving for $\mathbf{D}$ and taking only the diagonal (in accordance with the defini-

tion of factor analysis, where $\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ is full and $\mathbf{D}$ is diagonal):

$$
\begin{aligned}
\therefore \mathbf{D} &= \operatorname{diag}\left[\frac{1}{I}\sum_{i=1}^{I} \mathrm{E}_{\mathbf{w}}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] + \mu_{\mathbf{p}}\mathbf{y}^{\mathrm{T}} + \mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i{\mathbf{h}_i}^{\mathrm{T}}]\mathbf{\Phi}^{\mathrm{T}} - \right. \\
&\left. 2\mathrm{E}_{\mathbf{w}}[\mathbf{w}]\mathbf{y}^{\mathrm{T}} - 2\mathrm{E}_{\mathbf{w}}[\mathbf{w}]\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}} + 2\mu_{\mathbf{p}}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\right] \\
&= \operatorname{diag}\left[\frac{1}{I}\sum_{i=1}^{I} \mathbf{m}\mathbf{m}_{\mathbf{r}}{}^{\mathrm{T}} + \mathbf{L}_i^{-1} + \mu_{\mathbf{p}}\mathbf{y}^{\mathrm{T}} + \mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i{\mathbf{h}_i}^{\mathrm{T}}]\mathbf{\Phi}^{\mathrm{T}} - \right. \\
&\left. 2\mathbf{m}\mathbf{y}^{\mathrm{T}} - 2\mathbf{m}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}} + 2\mu_{\mathbf{p}}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\right] \\
&= \operatorname{diag}\left[\frac{1}{I}\sum_{i=1}^{I} (\mathbf{m}-\mu_{\mathbf{p}})(\mathbf{m}-\mu_{\mathbf{p}})^{\mathrm{T}} + \mathbf{L}_i^{-1} + \right. \\
&\left. \mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i{\mathbf{h}_i}^{\mathrm{T}}]\mathbf{\Phi}^{\mathrm{T}} - 2\mathbf{m}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}} + 2\mu_{\mathbf{p}}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\right] \\
&= \operatorname{diag}\left[\frac{1}{I}\sum_{i=1}^{I} (\mathbf{m}-\mu_{\mathbf{p}})(\mathbf{m}-\mu_{\mathbf{p}})^{\mathrm{T}} + \mathbf{L}_i^{-1} + \right. \\
&\left. (\mathbf{m}-\mu_{\mathbf{p}})\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}} - 2\mathbf{m}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}} + 2\mu_{\mathbf{p}}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\right] \\
&= \operatorname{diag}\left[\frac{1}{I}\sum_{i=1}^{I} (\mathbf{m}-\mu_{\mathbf{p}})(\mathbf{m}-\mu_{\mathbf{p}})^{\mathrm{T}} + \mathbf{L}_i^{-1} - (\mathbf{m}-\mu_{\mathbf{p}})\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i]^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\right] \\
&= \operatorname{diag}\left[\frac{1}{I}\sum_{i=1}^{I} (\mathbf{m}-\mu_{\mathbf{p}})(\mathbf{m}-\mu_{\mathbf{p}})^{\mathrm{T}} + \mathbf{L}_i^{-1} - \mathbf{\Phi}\mathrm{E}_{\mathbf{h}_i}[\mathbf{h}_i](\mathbf{m}-\mu_{\mathbf{p}})^{\mathrm{T}}\right] \quad \text{(F.44)}
\end{aligned}
$$

where $\mathrm{E}_{\mathbf{w}}[\mathbf{w}\mathbf{w}^{\mathrm{T}}] = \mathbf{m}\mathbf{m}_{\mathbf{r}}{}^{\mathrm{T}} + \mathbf{L}_i^{-1}$. □

References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[2] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech." in *Odyssey*, 2010, p. 6.

[3] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.

[4] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[5] S. J. Prince, *Computer vision: models, learning, and inference*. Cambridge University Press, 2012.

[6] Y. Jiang, K. A. Lee, and L. Wang, "PLDA in the i-supervector space for text-independent speaker verification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–13, 2014.

[7] J. A. Villalba and E. Lleida, "Handling i-vectors from different recording conditions using multi-channel simplified plda in speaker recognition." in *ICASSP*, 2013, pp. 6763–6767.

[8] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition." in *Interspeech*, 2006, pp. 1471–1474.

[9] P. Rajan, A. Afanasyev, V. Hautamäki, and T. Kinnunen, "From single to multiple enrollment i-vectors: Practical plda scoring variants for speaker verification," *Digital Signal Processing*, 2014.

[10] National Institute of Standards and Technology, "The NIST 2008 SRE Evaluation Plan," 2008. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2008/

[11] ——, "The NIST 2010 SRE Evaluation Plan," 2010. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/

[12] M. McLaren and D. Van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5456–5459.

[13] ——, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 755–766, 2012.

[14] N. Dehak, Z. N. Karam, D. A. Reynolds, R. Dehak, W. M. Campbell, and J. R. Glass, "A channel-blind system for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4536–4539.

[15] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[16] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[17] R. Raina, A. Y. Ng, and D. Koller, "Constructing informative priors using transfer learning," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 713–720.

[18] J. Yang, R. C. van Dalen, S.-X. Zhang, and M. Gales, "Infinite structured support vector machines for speech recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3320–3324.

[19] S. E. Shepstone, K. A. Lee, H. Li, Z.-H. Tan, and S. H. Jensen, "Source specific informative prior for i-vector extraction," in *Acoustics Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, Accepted Paper.

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

[21] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[22] A. Dempster, N. Laird, and D. B. Rubin, "Maximum likelihood estimation via the EM algorithm," *J. of the Stat. Roy. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[23] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[24] N. Brümmer, "The EM algorithm and minimum divergence," *Online http://niko. brummer. googlepages. Agnitio Labs Technical Report*, 2009.

[25] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L. R. Dai, "Minimum divergence estimation of speaker prior in multi-session PLDA scoring," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4007–4011.

[26] N. Brümmer, "EM for probabilistic LDA," 2010.

[27] P. Kenny, "A small foot-print i-vector extractor," in *Proc. Odyssey*, 2012.

[28] K. A. Lee, A. Larcher, C. H. You, B. Ma, and H. Li, "Multi-session PLDA scoring of i-vector for partially open-set speaker detection." in *INTERSPEECH*, 2013, pp. 3651–3655.

[29] Y. Jiang, K. A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, "PLDA modeling in i-vector and supervector space for speaker verification." in *INTERSPEECH*, 2012.

[30] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," DTIC Document, Tech. Rep., 1997.

Part III

Appendix A: Additional Experiments for Paper F

Additional Experiments for Paper F

1 Background

In Papers E [1] and F [2] we addressed the issue of source variability using source-specific informative priors. Unlike previous approaches that deal with source variability directly in the i-vector space, this approach operates in the supervector space, meaning that the proposed algorithm is certainly more computationally expensive than the previous i-vector based approaches. The purpose of this appendix is to add additional justification for our approach by comparing it to one of the previous state-of-the-art i-vector based algorithms.

In [3, 4] the authors present a proposal called *Source-Normalized-and-Weighted LDA* (SNAW). Two problems observed when using LDA on raw i-vectors are that the source variability negatively impacts the estimation of the between-speaker covariance matrix $\mathbf{S}_B$, and that insufficient utterances from each source for each speaker negatively effects the estimation of the within-speaker covariance matrix $\mathbf{S}_W$. To this end, the authors propose to source-normalize the i-vectors with respect to the source mean when computing the inter-speaker scatter matrix $\mathbf{S}_B$. As shown in [3, 4], for each source, assuming it has N_{src} utterances, the source mean is computed as

$$\mu_{\mathbf{src}} = \frac{1}{N_{src}} \sum_{n=1}^{N_{src}} \mathbf{w}_n \tag{45}$$

For the given source, assuming $\mu_{\mathbf{s}}$ is the mean-vector for each speaker from that source and N_s is the number of utterances for a speaker of a given source, the source-specific scatter matrix can be computed as:

$$\mathbf{S}_{B\,src} = \sum_{s=1}^{S_{src}} N_s(\mu_{\mathbf{s}} - \mu_{\mathbf{src}}).(\mu_{\mathbf{s}} - \mu_{\mathbf{src}})^{\mathrm{T}} \tag{46}$$

The overall weighted source-normalized scatter matrix across all source is given as:

$$\mathbf{S}_B = \sum \frac{N_{src}}{N} \mathbf{S}_{B\,src} \tag{47}$$

The within-speaker scatter matrix is then computed as the residual variability in the i-vector space:

$$\mathbf{S}_W = \mathbf{S}_T - \mathbf{S}_B \tag{48}$$

where $\mathbf{S}_T$ is computed as:

$$\mathbf{S}_T = \sum_{n=1}^{N} \mathbf{w}.\mathbf{w}^{\mathrm{T}} \tag{49}$$

2 Experimental Work: Additional Baseline

Using the same training data as that used in [1, 2], we incorporate an additional baseline. This is shown in Tables 3 and 4 as baseline A4, and corresponds to the standard LDA computation being replaced with the SNAW approach. For easy comparison, we include all previous experiment results. For the new baseline A4, the $\mathbf{T}$ matrix corresponding to the *pooled* configuration was used.

3 Discussion of Results for 2-prior i-vector Extraction

Looking at the SRE'08 results with respect to Equal Error Rate (EER), for female data, A4 was the best baseline for five out of eight common conditions (CCs), and for male data, it was the best for seven out of eight CCs. It challenged our proposal somewhat, since for the proposed configuration B1, when we include the additional baseline, we noticed that slightly fewer baselines were beaten than if A4 were not included. However, there was still good performance overall. For female trials, proposal B1 beat all four baselines for four out of eight CCs, and for male trials, for three out of eight CCs. B1 was also able to beat the A4 baseline (SNAW) for six out of eight CCs in the female case, and four out of eight CCs for the male case.

Looking at the SRE'10 results, again with respect to Equal Error Rate (EER) for female and male trials, A4 was the best baseline for four out of nine CCs. We saw markedly better performance for SRE'10 data than for SRE'08 data. For female trials, the configuration B1 beat the four baselines for six out of nine CCs, and for male trials, for five out of nine CCs. This was the same as reported for the three baseline case, which is testament to the strength of using a prior-based

Table 3: SRE'08 performance comparison for the eight common conditions (CC) in the short2-short3 core task. Left: FEMALE Trials, Right: MALE Trials. Each entry consists of EER (top) and DCF-08 (bottom). Entries in **bold** either match or beat the 4 baselines in Category A. Starred entries (*) in rows C1-C4 shown for cases that outperform B1. Plus entries (+) shown where D3 outperforms D1 and D2. NI is Non-informative prior (only 1).

			CC1: int-int		CC2: int-int		CC3: int-int		CC4: int-tel		CC5: tel-mic		CC6: tel-tel		CC7: tel-tel		CC8: tel-tel	
Configuration	T	I-vector	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
A1	Telephone-only	NI	3.51 0.16	2.84 0.13	1.50 0.05	0.32 0.00	3.61 0.17	2.97 0.13	5.69 0.28	4.07 0.19	6.65 0.27	4.17 0.19	5.85 0.31	4.67 0.25	2.73 0.14	2.32 0.11	3.24 0.15	1.43 0.06
A2	Pooled	NI	3.22 0.16	2.54 0.13	1.28 0.06	0.33 0.01	3.29 0.16	2.64 0.13	4.65 0.25	3.89 0.19	5.62 0.26	3.05 0.14	5.86 0.29	4.15 0.25	2.84 0.14	1.60 0.11	3.32 0.14	1.04 0.07
A3	Cascade	NI	3.17 0.16	3.01 0.14	1.25 0.05	0.41 0.02	3.27 0.16	3.22 0.15	5.38 0.26	4.27 0.20	6.10 0.26	4.12 0.16	5.86 0.30	4.06 0.23	2.98 0.13	1.66 0.10	3.81 0.15	1.32 0.07
A4	SNAW	NI	2.57 0.13	2.29 0.10	1.42 0.06	0.16 0.01	2.61 0.13	2.39 0.11	4.49 0.22	3.57 0.18	5.65 0.23	3.50 0.14	5.78 0.29	3.74 0.24	2.91 0.13	1.44 0.11	3.22 0.13	0.83 0.06
B1	Pooled	2-priors	**2.34** **0.12**	**1.95** **0.09**	1.32 0.06	0.32 0.00	**2.39** **0.12**	**2.04** **0.10**	**4.32** 0.23	3.91 0.20	**5.37** 0.26	3.21 0.17	5.79 **0.29**	3.84 0.24	2.87 **0.13**	**1.39** 0.11	3.27 **0.12**	0.90 0.07
C1	2-priors	2-priors	**2.33*** 0.12	**1.96** **0.10**	**1.20*** **0.05***	0.24* 0.00	**2.37*** **0.12**	**2.04** **0.10**	**4.24*** **0.22***	3.90* 0.19*	**5.07*** **0.23***	3.09* 0.15*	**5.72*** **0.28***	**3.67*** 0.24	2.80* **0.13**	**1.42** 0.11	3.29 **0.13**	0.87* 0.07
C2	2-priors	2-priors	**2.26*** 0.12	**1.99** **0.10**	1.58 0.06	0.18* 0.01	**2.28*** **0.13**	**2.09** **0.11**	**4.20*** **0.21***	3.94 0.20	**4.83*** **0.23***	**2.97*** 0.16*	**5.68*** **0.29**	3.82* 0.24	2.74* **0.13**	1.56 0.11	3.42 0.14	**0.76*** 0.07
C3	2-priors	2-priors	**2.25*** 0.12	**1.99** **0.10**	1.56* 0.06	0.27* **0.00**	**2.29*** **0.12**	**2.09** **0.10**	**4.24*** **0.21***	3.86* 0.20	**5.09*** **0.23***	3.08* 0.15*	**5.68*** **0.29**	3.94 0.24	2.88 **0.13**	**1.47** 0.10	3.41 **0.13**	**0.81*** **0.06**
C4	2-priors	2-priors	**2.40** **0.13**	**2.01** **0.10**	1.60 0.06	0.27* **0.00**	**2.44** **0.13**	**2.10** **0.10**	**4.05*** **0.21***	3.98 0.20	**5.01*** 0.24*	3.06* 0.15*	**5.70*** **0.28***	3.79* 0.24	2.84* **0.13**	**1.35*** 0.11	3.34 0.14	**0.78*** 0.07
D1	2-priors	3-priors	4.57 0.20	3.18 0.15	2.50 0.06	0.41 0.01	4.69 0.21	3.33 0.16	7.33 0.37	4.59 0.24	5.15 0.24	3.24 0.14	5.67 0.29	3.72 0.23	2.83 0.14	1.43 0.10	3.26 0.14	0.66 0.07
D2	2-priors	3-priors	3.40 0.16	2.67 0.13	1.99 0.04	0.41 0.02	3.48 0.17	2.78 0.14	6.28 0.34	4.14 0.21	5.06 0.23	3.26 0.14	5.69 0.29	3.80 0.24	2.80 0.13	1.50 0.11	3.44 0.14	0.85 0.07
D3	2-priors	3-priors	**3.00+** **0.15+**	2.52+ 0.13	1.73+ **0.04**	0.27+ 0.03	3.07+ 0.16+	2.64+ 0.14	5.77+ 0.31+	4.18 0.20+	**5.05+** 0.24	3.27 0.15	**5.71** **0.28+**	3.95 0.25	2.83 **0.13**	1.52 0.11	3.46 **0.13+**	**0.73+** 0.07

Table 4: SRE'10 performance comparison for the nine common conditions (CC) for the core-core core task. Left: FEMALE Trials, Right: MALE Trials. Each entry consists of EER (top) and DCF-10 (bottom). Entries in **bold** either match or beat the 4 baselines in Category A. Starred entries (*) in rows C1-C4 shown for cases that outperform B1. Plus entries (+) shown where D3 outperforms D1 and D2. NI is Non-informative prior (only 1).

			CC1: int-int-same-mic		CC2: int-int-diff-mic		CC3: int-tel		CC4: int-mic		CC5: nve-nve-diff-tel		CC6: nve-hve-diff-tel		CC7: nve-hve-mic		CC8: nve-lve-diff-tel		CC9: nve-lve-mic	
Configuration	T	I-vector	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M
A1	Telephone-only	NI	3.06 0.36	2.02 0.22	5.65 0.67	3.45 0.47	4.21 0.57	3.55 0.50	3.96 0.54	2.72 0.36	3.59 0.40	3.47 0.46	8.09 0.79	4.64 0.70	8.49 0.69	4.91 0.46	2.01 0.43	1.14 0.23	2.46 0.27	1.54 0.22
A2	Pooled	NI	3.16 0.34	2.22 0.20	5.13 0.60	3.14 0.40	3.34 0.55	2.82 0.40	3.78 0.50	2.54 0.36	3.00 0.37	2.60 0.45	7.13 0.79	4.01 0.67	7.98 0.70	4.95 0.45	1.66 0.29	1.54 0.23	2.55 0.23	1.34 0.24
A3	Cascade	NI	3.12 0.38	2.29 0.24	5.60 0.61	3.29 0.43	4.01 0.52	2.62 0.49	4.04 0.51	2.87 0.33	3.41 0.34	3.13 0.36	7.10 0.82	4.33 0.74	8.19 0.74	5.25 0.45	1.83 0.29	1.68 0.26	3.08 0.30	1.61 0.21
A4	SNAW	NI	2.82 0.34	2.18 0.22	4.92 0.57	2.95 0.40	3.76 0.57	2.80 0.47	3.34 0.50	2.26 0.33	3.23 0.40	2.88 0.36	6.93 0.80	4.47 0.66	8.27 0.66	4.27 0.49	1.67 0.40	1.84 0.32	2.83 0.29	1.41 0.20
B1	Pooled	2-priors	**2.43** 0.35	**1.67** 0.24	**4.44** 0.62	**2.25** 0.43	3.87 0.62	3.19 0.61	**3.33** **0.49**	**2.22** 0.36	**3.00** 0.42	2.89 0.39	7.11 **0.77**	4.13 0.69	**7.49** **0.59**	**4.16** 0.51	**1.59** 0.34	1.56 0.26	2.48 0.31	**1.15** 0.24
C1	2-priors	2-priors	**2.56** **0.24***	**1.48*** 0.24	**4.54** **0.45***	**2.20*** 0.45	3.86* 0.60*	2.87* 0.59*	**3.32*** **0.38***	**2.12*** 0.38	3.13 0.40*	2.82* 0.40	7.19 0.67*	4.39 **0.67***	**7.69** **0.48***	4.67 **0.48***	**1.65** **0.23***	1.51* **0.23***	2.65 0.25*	**0.85*** 0.26
C2	2-priors	2-priors	**2.67** 0.37	**1.69** 0.26	**4.62** **0.57***	**2.54** **0.39***	3.75* 0.60*	3.07* 0.58*	**3.28*** **0.49**	**2.24** 0.36	3.26 0.41*	2.76* 0.38*	7.18 0.80	4.18 **0.66***	**7.57** **0.59**	4.61 **0.42***	**1.61** 0.36	1.47* **0.18***	2.68 0.27*	**1.16** 0.28
C3	2-priors	2-priors	**2.63** 0.36	**1.81** 0.24	**4.62** 0.58*	**2.58** 0.42*	3.83* 0.63	3.02* 0.52*	**3.39** **0.49**	2.27 **0.33***	3.26 0.40*	2.78* 0.40	**7.08*** **0.75***	4.10* **0.64***	**7.37*** **0.57***	4.72 **0.42***	**1.67** 0.38	1.34* **0.18***	2.70 0.27*	**1.34** **0.21***
C4	2-priors	2-priors	**2.61** 0.36	**1.79** 0.24	**4.60** **0.56***	**2.53** 0.42*	3.79* 0.60*	3.09* 0.57*	**3.29*** 0.53	2.30 0.34*	3.12 0.42	2.90 0.39	7.22 0.80	4.01* 0.69	**7.81** **0.58***	4.48 **0.44***	**1.62** 0.38	1.62 **0.18***	2.64 0.27*	**1.27** 0.26
D1	2-priors	3-priors	4.48 0.48	2.40 0.34	8.45 0.78	3.41 0.53	6.87 0.74	4.12 0.66	4.98 0.66	2.87 0.38	3.03 0.41	2.88 0.41	7.32 0.74	4.09 **0.63**	**7.71** **0.67**	**4.15** **0.42**	1.67 0.39	1.64 **0.19**	2.61 0.29	**0.89** 0.24
D2	2-priors	3-priors	3.82 0.38	2.05 0.27	6.58 0.74	**3.01** 0.49	6.23 0.75	4.14 0.69	4.74 0.66	3.00 0.43	3.23 0.41	2.62 0.41	**6.98** 0.80	4.19 0.70	8.12 **0.66**	4.55 **0.39**	**1.63** 0.38	1.65 **0.22**	2.71 0.27	1.37 0.25
D3	2-priors	3-priors	3.60+ 0.37+	2.07 0.28	6.11+ 0.69+	**2.95+** 0.48+	5.69+ 0.70+	4.02+ 0.61+	4.56+ 0.63+	2.76+ 0.40	3.20 0.39+	2.73 0.40+	**6.69+** **0.79**	**3.95+** 0.69	**7.90** **0.64+**	4.46 **0.39**	**1.63** 0.37+	1.64 **0.18+**	2.54+ 0.26+	**1.29** 0.26

approach. Comparing B1 directly with A4, B1 was able to beat A4 for seven out of nine CCs for both female and male trials.

4 Conclusion

We added the SNAW algorithm for computing the LDA transform in the i-vector space as an additional baseline. The addition of this new baseline resulted in fewer common conditions having a better EER when compared to all four baselines for the SRE'08 case and no change for the SRE'10 case. 2-prior i-vector extraction was able to beat the SNAW LDA approach for a substantial number of CCs.

References

[1] S. E. Shepstone, K. A. Lee, H. Li, Z.-H. Tan, and S. H. Jensen, "Source specific informative prior for i-vector extraction," in *Acoustics Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, Accepted Paper.

[2] ——, "Total variability modeling using source-specific priors," *Audio, Speech, and Language Processing, IEEE Transactions on*, Submitted Paper.

[3] M. McLaren and D. Van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5456–5459.

[4] ——, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 755–766, 2012.