ROBUST MACHINE LEARNING METHODS FOR COMPUTATIONAL PARALINGUISTICS AND MULTIMODAL AFFECTIVE COMPUTING

by

Heysem Kaya

B.S., Computer Education and Educational Technology, Boğaziçi University, 2006

M.S., Computer Engineering, Bahçeşehir University, 2009

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Computer Engineering PhD Program

Boğaziçi University

2015

ACKNOWLEDGEMENTS

This thesis is dedicated to my wife, Natalia, who has been patiently waiting for the completion of my PhD study. Without her love, support and understanding, this thesis could not reach its current state. My gratitude to all members of my huge family for their support and motivation.

With gratitude to my advisor Prof. Albert Ali Salah for his invaluable technical support and guidance.

With gratitude to my co-advisor Prof. Fikret Gürgen for his patience and support in various problems encountered during the thesis study.

With gratitude to Prof. Lale Akarun for her guidance at the very beginning of my PhD quest and her support during my endeavor.

With gratitude to my thesis supervisory committee members Prof. Taylan Cemgil and Prof. Çiğdem Eroğlu Erdem, for their constructive feedback and fruitful discussions.

With gratitude to Prof. Björn Schuller and Prof. Alexey Karpov, with whom I had the pleasure of collaboration during my visit to their institutions. With gratitude to Prof. Ethem Alpaydın who has matured my machine learning understanding and to Prof. Olcay Kurşun for encouraging me in pursuing this PhD study.

I would like to thank my colleagues Dr. Sibel Halfon, Tuğçe Özkaptan, Şefika Yüzsever, Furkan Gürpınar, Neşe Alyüz, Umut Konur, Furkan Kıraç, Pınar Sağlam, Yunus Emre Kara, Gaye Genç, Gül Varol, and Özlem Salehi for the research motivation they provided. Lastly, I appreciate and thank TÜBİTAK for the financial support provided via BİDEB 2211 programme.

ABSTRACT

ROBUST MACHINE LEARNING METHODS FOR COMPUTATIONAL PARALINGUISTICS AND MULTIMODAL AFFECTIVE COMPUTING

The analysis of affect (e.g. emotions or mood), traits (e.g. personality), and social signals (e.g. frustration, disagreement) are of increasing interest in human computer interaction, in order to drive human-machine communication to become closer to human-human communication. It has manifold applications ranging from intelligent tutoring systems to affect sensitive robots, from smart call centers to patient telemonitoring. The study of computational paralinguistics, which covers the analysis of speaker states and traits, faces with real life challenges of inter-speaker and inter-corpus variability. In this thesis, machine learning methods addressing these challenges are targeted. Automatic model selection methods are explored for modeling high dimensional paralinguistics data. These approaches can deal with different sources of variability in a parametric manner. To provide statistical models and classifiers with a compact set of potent features, novel feature selection methods based on discriminative projections are introduced. In addition, multimodal fusion techniques are sought for robust affective computing in the wild. The proposed methods and approaches are validated over a set of recent challenge corpora, including INTERSPEECH Computational Paralinguistics Challenge (2013-2015), Audio-Visual Emotion Challenge (2013/2014), and Emotion Recognition in the Wild Challenge 2014. The methods proposed in this thesis advance the state-of-the-art in most of these corpora and yield competitive results in others, while enjoying the properties of parsimony and computational efficiency.

ÖZET

HESAPLAMASAL PARALİNGUİSTİK ve ÇOK-KİPLİ DUYUŞSAL HESAPLAMA İÇİN GÜRBÜZ YAPAY ÖĞRENME YÖNTEMLERİ

İnsan-makina iletişimini insan-insan iletişimine yaklaştırmak için, insan-makina etkileşim alanında duygu durum (örn. duygu, ruh hali), özellik (örn. kişilik) ve sosyal işaretler (örn. düş kırıklığı, fikir ayrılığı) analizine artan ilgi söz konusudur. Bunun akıllı eğitim sistemlerinden duyguları anlayabilen robotlara, akıllı çağrı merkezlerinden uzaktan hastaları takip eden sistemlere kadar çeşitli uygulamaları vardır. Konuşmacı durum ve özelliklerini kapsayan hesaplamasal paralinguistik çalışma alanı, konuşmacı ve veritabanı değişkenliği gibi gerçek hayat problemleriyle yüzleşmektedir. Bu tezde, bu problemleri çözmek için çeşitli yapay öğrenme yöntemleri geliştirilmesi hedeflenmiştir. Yüksek boyutlu paralinguistik verilerin modellenmesi için otomatik model seçim yöntemleri geliştirilmiştir. Bu yaklaşımlar farklı değişkenlik kaynaklarını parametrik bir şekilde ele alabilmektedir. İstatistiksel modeller ve sınıflayıcılara özlü, potansiyeli yüksek öznitelikler sağlamak için ayrımsayıcı izdüşüm tabanlı yeni değişken seçim yöntemleri tanıtılmıştır. Ek olarak, zorlu koşullarda gürbüz duyuşsal hesaplama için çok-kipli tümleştirme teknikleri irdelenmiştir. Önerilen yöntem ve yaklaşımlar INTERSPEECH Computational Paralinguistics Challenge (2013-2015), Audio-Visual Emotion Challenge (2013/2014), ve Emotion Recognition in the Wild Challenge 2014 gibi bir dizi yakın tarihli yarışma veri kümelerinde geçerlenmiştir. Bu tezde önerilen yöntemler sadelik ve hesaplamasal verimlilik özelliklerini taşımakla beraber, bu veri kümelerinin çoğunda problem üzerinde raporlanmış en iyi çözümlere çok yakın veya daha yüksek başarı elde etmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\mathbf{C_{XY}}$	Cross-set covariance between sets $\boldsymbol{X}$ and $\boldsymbol{Y}$
$F0$	Fundamental Frequency
$\mathbf{H}$	Hidden Output Matrix
$\boldsymbol{T}$	Label Matrix
$\boldsymbol{w}$	Weight Vector
$\boldsymbol{W}$	Mapping Matrix
$\boldsymbol{x}$	Random Variable
$\boldsymbol{X}$	Feature set
$\mathcal{X}$	Dataset
$\boldsymbol{Y}$	Target Variable
$\boldsymbol{z}$	Latent Variable
β	ELM second layer projection matrix
$\boldsymbol{I}$	Identity Matrix
λ	Eigenvalue
$\mathbf{\Lambda}$	Factor Loading Matrix
$\boldsymbol{\mu}$	Mean
$\mathcal{G}_k$	Mixture Component
$\mathbf{\Sigma}$	Covariance Matrix
$\mathbf{\Psi}$	Diagonal Uniquenesses Matrix
ρ	Correlation Coefficient
τ	SVM complexity / Kernel ELM regularization parameter
κ	Smoothing Parameter

LIST OF ACRONYMS/ABBREVIATIONS

AFEW	Acted Facial Expressions in the Wild
AMoFA	Adaptive Mixture of Factor Analyzers
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ASC	Affect Sub-challenge
ASR	Automatic Speech Recognition
AUC	Area Under Precision-Recall Curve
AVEC	Audio-Visual Emotion Challenge
BIC	Bayesian Information Criterion (Schwarz Criterion)
BP	Back Propagation
BUEMODB	Boğaziçi University Emotional Speech Database
CCA	Canonical Correlation Analysis
CFS	Correlation-based Feature Selection
CNN	Convolutional Neural Network
ComParE	Computational Paralinguistics Challenge
CORR	Pearson's Correlation
CV	Cross Validation
DES	Danish Emotional Speech Database
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DP	Dirichlet Process
DPMM	Dirichlet Process Mixture Model
DSC	Depression Sub-challenge
EC	Eating Condition
EF	Equal Weight Fusion
ELM	Extreme Learning Machine
EM	Expectation-Maximization
EMODB	Berlin Emotional Speech Database

ENTERFACE	ENTERFACE Emotional Speech Database
EW	Equal Width
FA	Factor Analysis
FDA	Fisher Discriminant Analysis
FS	Feature Selection
GA	Genetic Algorithm
GMM	Gaussian Mixture Model
GP	Gaussian Process
HAC	Hierarchical Agglomerative Clustering
HNR	Harmonics-to-Noise Ratio
HSD	Honest Significant Difference
IMM	Infinite Mixture Model
IMoFA	Incremental Mixture of Factor Analyzers
JFA	Joint Factor Analysis
KNN	K-Nearest Neighbors
LBP	Local Binary Patterns
LBP-TOP	Local Binary Patterns from Three Orthogonal Planes
LDA	Linear Discriminant Analysis
LFDA	Local Fisher Discriminant Analysis
LGBP	Local Gabor Binary Patterns
LGBP-TOP	Local Gabor Binary Patterns from Three Orthogonal Planes
LLD	Low Level Descriptor
LPQ	Local Phase Quantization
LPP	Locality Preserving Projections
LSSVM	Least Square Support Vector Machines
MAE	Mean Absolute Error
MDL	Minimum Description Length
MFCC	Mel-Frequency Cepstral Coefficient
MI	Mutual Information
MML	Minimum Message Length
MoFA	Mixture of Factor Analyzers

MoG	Mixture of Gaussians
MPCCA	Mixtures of Probabilistic CCA
mRMR	Minimum Redundancy and Maximum Relevance
MSE	Mean Square Error
NID	Normalized Information Distance
UAR	Unweighted Average Recall
UAAUC	Unweighted Average Area Under Precision Recall Curve
ULFMM	Unsupervised Learning of Finite Mixture Models
PCA	Principal Component Analysis
PCC	Pearson's Correlation Coefficient
PCCA	Probabilistic CCA
PLDA	Probabilistic LDA
PLP	Perceptual Linear Prediction
PPCA	Probabilistic Principal Component Analysis
RASTA	Relative Spectra
RASTA-PLP	RASTA Style Perceptual Linear Prediction
RBF	Radial Basis Function
RF	Random Forests
RMSE	Root Mean Square Error
SER	Speech Emotion Recognition
SHS	Subharmonic Summation
SLFN	Single Layer Feed-forward Network
SVM	Support Vector Machine
TOP	Three Orthogonal Planes
VB	Variational Bayes
VBMoFA	Variational Bayesian Mixture of Factor Analyzers
WD-KNN	Weighted Discrete K-Nearest Neighbors
WF	Weighted Fusion

1. INTRODUCTION

Non-verbal communication constitutes an important part of all human-human communications. Facial expressions, hand and body gestures, posture, gaze, and most importantly, the speech prosody are used in non-verbal communication. These signals can have meanings on their own, or can be used to modulate or change verbal communication. Some of these are voluntary, whereas some are not. For instance, it is possible to communicate fatigue through the facial expression or through paralinguistic qualities of speech.

In recent years, research in affective computing, human-computer interaction and speech processing focused on computational paralinguistics, which deals with automatic analysis of the traits (e.g. cultural background, personality) and the states (e.g. sleepiness, emotions) of the interacting individual. Moreover, computer based systems can also be employed to detect and monitor the level of a long term illness, such as depression, which is inherently related to emotions.

The study of computational paralinguistics deals with automatic analysis of speaker's states and traits rather than the spoken content [1, 2]. One major challenge in the field for a real-life application is handling the large variability stemming from speakers, acoustic recording conditions, spoken content, used language, culture, and such [3]. In this regard, the main research problem of this thesis is finding efficient and robust speaker-independent methods for affect recognition/prediction in computational paralinguistics.

The state-of-the-art computational paralinguistics applications are built using suprasegmental features obtained from functionals (e. g. moments, extremes) operating on frame-level Low Level Descriptors (LLD) e. g. $F0$, MFCC, jitter, shimmer [1]. Brute-force extraction of high-dimensional potent features is commonly encountered in competitive baseline feature sets of the most recent computational paralinguistics challenges [4–6]. One drawback of feature brute-forcing is the curse of dimensional-

ity, which is defined as limited coverage of input space with low number of instances as opposed to massive dimensionality. Furthermore, such an extensive and large feature set usually contains redundant and irrelevant features that impede generalization. Therefore, one research direction of this thesis is feature selection (FS).

In multivariate statistics and pattern recognition, mixture models are intended to cope with large variability, because they model the data as a combination (i.e. the mixture) of local distributions [7]. While such a mixture gives modeling flexibility, the number of components to use in the mixture, as well as the shape of each component should be carefully selected. In this thesis, one of the goals was to develop an efficient and accurate automatic mixture model selection method.

All of the mentioned issues are related to the main research question of this thesis: "What are efficient, robust, speaker-independent methods for computational paralinguistics and multimodal affective computing?". The research sub-problems are: (i) "How can we improve robustness of paralinguistic/affective computing systems with efficient feature selection?", (ii) "What are efficient and robust model learning approaches that can be practically used in the field?" (iii) "How can we benefit from multimodality for improving robustness?", and (iv) "What are alternative utterance representation methods that are more robust compared to the state-of-the-art approach?".

Automatic model selection for MoFA. To address the mixture model selection issue, a fast and parsimonious model selection algorithm called *Adaptive Mixture of Factor Analyzers* (AMoFA) [8] is proposed. In MoFA, each component's covariance structure is compactly modeled via Factor Analysis (FA). The proposed method is capable of adapting a mixture model to data by selecting an appropriate number of components and factors per component. AMoFA uses a Minimum Message Length (MML) based criterion to control the model complexity, and is also capable of removing factors and components when necessary. The efficacy of AMoFA algorithm is evaluated in clustering, manifold learning, as well as class conditional modeling (classification).

Novel feature selection methods. Despite its efficiency and accuracy in modeling high dimensional data, AMoFA also requires a compact set of relevant features. For this purpose, as well as to improve discriminative models, we have investigated several FS methods in this thesis. Novel discriminative projection based FS methods were proposed with successful application to acoustic prediction of (i) depression severity level [9], (ii) physical load [10] and (iii) conflict [11].

Applying recent machine learning paradigms to affective computing. We have applied state-of-the-art machine learning approaches to the problems of computational paralinguistics and affective computing. We have contrasted Partial Least Squares (PLS) [12] based classifier and Extreme Learning Machines (ELM) [13, 14]. These methods learn much faster than Support Vector Machines (SVM) that are extensively used in the literature, and generalize well to unseen data. We have also explored decision level fusion of these base classifiers, and applied them to recent audio-visual challenges [15, 16], where the tasks were *emotion recognition in-the-wild* [17, 18] and prediction of depression severity level [19].

Utterance representation. We have investigated alternative utterance representation and normalization schemes to alleviate the speaker variability. Fisher vector encoding of descriptors are recently popular in computer vision and are shown to provide state-of-the-art performance in image retrieval and activity recognition [20, 21]. In our framework, the FV is followed by cascaded normalization that is composed of speaker level z-normalization and non-linear normalization.

To sum up, the primary contributions of this thesis are a new set of discriminative projection based feature filters and a novel automatic mixture model selection method. The secondary contributions include employing and combining ELMs in audio-visual affective computing, and using the FV encoding with introduced cascaded normalization for paralinguistic analysis. The proposed methods are experimentally validated in recent challenge corpora adhering to standard protocols, achieving the state-of-the-art

results on many of them. A summary of proposed/adapted methods over the problems dealt with in this thesis are given in Table 1.1.

Table 1.1. Summary of thesis contributions over applications. **Bold** denotes novel methods proposed in this thesis, whereas *italic* indicates existing methods applied for the first time on the respective problem. A: Audio, V: Video. Details of existing methods are given in Chapter 2. Newly proposed methods are presented in Chapter 3.

Application	Feature Extraction	Feature Selection	Model Learning
Conflict (A: [11])	openSMILE	**SLCCA-RAND**	SVM
Eating Condition (A: [22])	*FV Encoding*	**SLCCA-RAND, SLCCA-LLD**	*ELM, PLS*
Emotion (A: [8, 23], V: [19], A/V: [17, 18])	A: openSMILE; V: LPQ, LBP-TOP, LGBP-TOP, Riemannian Kernels	A: **SLCCA-LLD**; V: Facial Region Selection	A:SVM, **AMoFA**; V: *ELM*, PLS
Depression (A: [9], A/V: [19, 24])	A: openSMILE, V: LPQ, LGBP-TOP, *CCA covariates*	A: **SLCCA-Filter, mRMR-CCA, MCR-CCA**	*TB, ELM*
Laughter (A: [25])	openSMILE	*mRMR*	*RF*
Personality Traits (A: [26])	openSMILE	**SLCCA-Filter**	*ELM*
Physical Load (A: [10])	openSMILE	**SLCCA-LLD**	SVM

The remainder of this thesis is organized as follows. In the next chapter, background on computational paralinguistics and the existing methods used in the thesis are presented. In Chapter 3, proposed discriminative projection based feature filters and the AMoFA algorithm are introduced. Applications of the proposed methods in paralinguistics and multimodal affective computing are given in Chapters 4 and 5, respectively. Finally, Chapter 6 concludes with a general discussion and future directions.

2. BACKGROUND

In this chapter, we provide background on Computational Paralinguistics (Section 2.1) and the existing methods used throughout the thesis (Section 2.2).

2.1. Computational Paralinguistics

Paralinguistics is the study of non-verbal communication that conveys emotion and nuances meaning. It deals with how the words are spoken rather than what is spoken. Speech Emotion Recognition (SER) is the branch of study that is categorized under paralinguistics, which is the broader study of Speaker State and Trait Recognition (SSTR). Speaker *states* correspond to conditions changing over time such as emotions, health state, interests, fatigue and stress, while speaker *traits* correspond to permanent or relatively permanent speaker characteristics such as personality, ethnicity, physical appearance and gender.

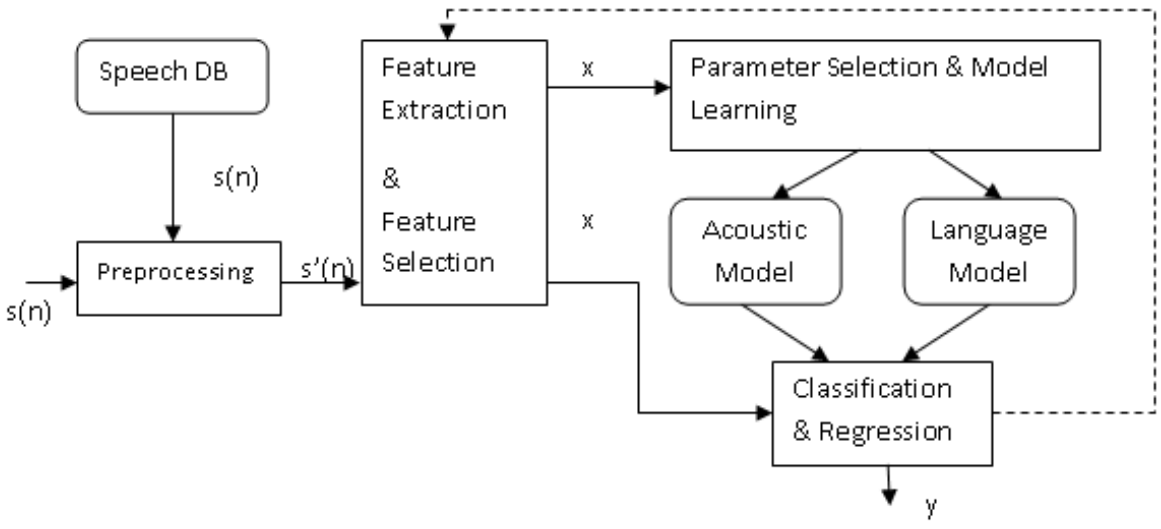

Figure 2.1. Speaker state and trait recognition pipeline. Adapted from Schuller [1].

The processing of speech for SSTR given by [1] is shown in Figure 2.1. The pipeline contains signal processing and pattern recognition subtasks. In audio-visual affective computing, the audio DB is replaced with video DB, preprocessing includes

face detection/alignment followed by visual feature extraction for corresponding modeling. It has been shown that when the automatic speech recognition (ASR) is put into the emotion recognition process, the language model improves the overall recognition given that the ASR is accurate. However, this is not always possible since the recognition of affective speech is itself a challenge [27]. We next briefly describe the elements of the pipeline.

Speech database contains speech audio files for model learning and testing. Also, it may comprise the representation of the spoken content and targets such as speaker emotion, age and personality. It is preferable that speech is recorded naturally and the number of speakers is as large as possible and the categorization of targets is reasonable and meaningful.

Pre-processing copes with the enhancement of signal properties of the speech. Speech can be consisted of multiple speakers and noise, therefore pre-processing is essential for the improvement of the speech quality. Pre-processing step is also employed after feature extraction, e. g. via normalization and up/down-sampling. *Feature extraction* is the phase where acoustic and linguistic features are extracted. This extraction depends on the problem of the research area. Feature extraction process also corresponds to transformation of original feature set to another space where much lower dimensions than the original dimensionality suffice for further pattern recognition task. Two well known and widely employed feature extraction techniques are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Both of them provide linear feature extraction however PCA is unsupervised, while LDA is supervised. Other two well known linear feature reduction methods, namely Multidimensional Scaling (MDS) and Factor Analysis (FA), have common characteristics with PCA [7, 28].

Feature Selection refers to finding the most significant features for the task at hand. This step is a challenging research area of the speech analysis. Feature extraction phase produces an extensive number of features, however all these may not be relevant to our task. Also the feature set may contain irrelevant and redundant features. Our

aim is to find an optimal feature set to improve generalization of the learner.

Classification/Regression deals with the categorization of the test data into either discrete or continuous targets. Classification process determines the targets such as emotion classes (anger, disgust, fear, happiness, sadness and surprise). While, regression process deals with prediction of continuous value of the targets such as speaker's height in cm, age in years. *Model learning* task is performed by the classifier/regressor, which trains the data and learns knowledge from the targets of data. *Acoustic Model* models learn dependencies between the acoustic features and the classes, or continuous values in the case of regression. *Language Model* models learned dependencies between linguistic features of the speech and the related targets.

Parameter Selection refers to optimizing the parameters of the learner model. During parameter selection process, speech instances for testing should not be utilized to avoid overestimation. In other words, for a proper treatment in terms of machine learning, the data should be split into three speaker disjoint sets, namely for training, development and testing. The training set is used to learn model parameters given features and hyper-parameters whereas the development set is used to optimize features and model hyper-parameters. The final "optimized" system is trained from the combination of training and the development sets is applied on testing to evaluate model performance in real-life conditions.

The challenges of handling emotional speech include but not limited to personal, cultural and environmental differences as well as weak cues of emotion in daily life speech. The databases are generally collected in isolated environments and the speech is sometimes "grotesque" i.e. over exaggerated. On the other hand, the study in this field is fruitful due to two reasons (i) Machine learning/pattern recognition methods are not well employed in the field, so state-of-the-art machine learning models will have significant contribution (ii) Although there are well suited features there is yet no feature set which is good for all emotion databases. The state-of-the-art in general paralinguistics tasks however, uses large scale suprasegmental feature extraction via passing a set of summarizing statistical functionals over the Low Level Descriptor

(LLD) contours. We continue with the review of features in the next section.

2.1.1. Review of Common Features and Classifiers

The processing of speech is followed by tasks of machine learning similar to other fields. In the this section we summarize the features and methods commonly used in SER. While features define "what to learn", the classification and regression methods determine "how" to learn the mapping function between the features and the targets.

2.1.2. Common Features

In computational paralinguistics field including SER, Pitch (also referred to as Fundamental Frequency-$F0$), Formants (resonant frequencies of vocal tract filter), Mel Frequency Cepstral Coefficients (MFCC), Modulation Spectrum, Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) [29, 30], Energy and variation features (i.e. Shimmer and Jitter which are usually employed in biomedical health care applications) are frequently used Low Level Descriptors (LLD). While among them MFCC and RASTA-PLP are the most widely used, Line Spectral Pair Frequency (LSP) [31] features are also successfully applied to SER [32, 33].

Pitch is defined as perceived level of $F0$, which is the frequency associated with vibration of the vocal cords. This prosodic feature provides important affect related cues. Signal energy and acoustic intensity are yet other commonly used features in SSTR. Jitter and shimmer are defined as micro perturbation of pitch and energy, respectively. Besides affect recognition, these variation features are commonly used for clinical telemonitoring of long term disorders such as depression and Parkinson's disease.

The most popularly used LLD in speech technologies, namely MFCC features correspond to Inverse Fourier Transform (IFT) or preferably the Discrete Cosine Transform (DCT) of the Mel-scaled log-magnitude spectrum of the signal. Mel-Scale mimics the human hearing capabilities in the way that it allows discriminating lower frequen-

cies better than the higher frequencies. Currently, MFCC are successfully applied to ASR and SSTR.

LSP feature representation of speech [31] provides efficient and robust estimation of formant frequencies. Formant frequencies, especially the first two, are known to carry affect related information [34].

While LLDs define the first level features obtained from the speech signal, further feature extraction is done via *functionals*. A functional is a function applied to other functions or distributions e. g. min, max, mean, mode, range of a distribution. The functionals project the LLD contours (over the speech analysis frames) to a fixed size vector. This is done for further pattern recognition process with fixed length features. It is also possible to use a variable length representation taking also into account the sequence information using HMM variants. HMMs have long been employed in ASR [35], and have been successfully used in SER [36]. However, the state-of-the-art results are not obtained with generative modeling using HMMs but using discriminative classifiers (e. g. SVMs) with utterance level features obtained via functionals. This is partly due to suprasegmental nature of the underlying phenomena and the advantage of discriminative learning. Note that not only raw LLDs, but also the first and second order derivatives are popularly used in speech recognition and in paralinguistic analysis [4,5].

Brute-forced extraction of suprasegmental features usually via the open source openSMILE tool [37] leads to a massive dimensionality. Despite the success of the approach without feature optimization, in order to tailor the feature set for the task at hand, dimensionality reduction methods are sought. In the literature, well known feature reduction methods such as PCA, LDA and Forward Selection based on CV are used in SER to avoid overfitting [34]. This thesis proposes CCA based filters [9–11].

2.1.3. Common Classifiers

The most commonly employed classifiers in SER are Gaussian Mixture Models (GMM), Artificial Neural Networks (ANN), Support Vector Machines (SVM) as well

as Hidden Markov Models (HMM) [34]. The state-of-the-art models of SER for the current databases are those trained with SVMs and Deep Neural Networks (DNN) [1]. DNNs family also comprise Convolutional Neural Networks (CNN) and Recurrent Networks (RNN) as that are popularly used in the state-of-the-art recognition systems. For details on CNNs and RNNs, the reader is referred to [7, 28]. From ANN family, Extreme Learning Machines (ELM), which combines fast model learning with accurate prediction capability, is recently introduced for multimodal emotion recognition in the wild [17, 18].

In a 2011 review paper, El Ayadi *et al.* [34] stress that some methods well established in machine learning are not thoroughly employed in SER. Combining multiple learners and multimodal late fusion are some examples. In affective computing literature, learner combination has been shown to outperform base learners (c. f. [38–40]). In this thesis, we successfully applied learner and modality fusion to laughter detection [25], depression severity level prediction [19, 24], multimodal emotion recognition [17, 18], and eating condition recognition [22].

2.1.4. Review of Affective Databases

In a review of the field, Ververidis and Kotropoulos [41] listed 32 affective databases for 10 languages as well as a multi-lingual database. Considering the statistics presented therein and recent corpora, we observe that the most frequently occurring emotions in collected DBs are Anger, Sadness, Fear and Happiness (see Figure 2.2 for details).

As stated before, one of the challenges is the naturalness of the collected data. Out of the 32 databases listed in [41]; 21 are acted, 8 are natural, in 2 half of the recordings are natural and half acted and one of them is semi-natural. We give a list of databases collected from recent reviews and challenges in Tables 2.1 and 2.2.

Table 2.1. Affective speech corpora.

Corpus	Language	States /Traits
AFEW [42]	English	Neutral, anger,disgust, fear, happiness, sadness, surprise
Amir *et al.* [43]	Hebrew	Anger, disgust, fear, joy, neutral, sadness
AVEC 2013 [5]	German	Continuous depression, valence and activation
AVEC 2014 [16]	German	Continuous depression, valence, dominance and activation
AVIC [44]	English	Three levels of interest (interested, netural, bored)
BabyEars [45]	English	Approval, attention, prohibition
BHUDES [46]	Mandarin	Anger, joy, sadness, fear, disgust, surprise
BUEMODB [23]	Turkish	Neutral, anger, sadness, happiness
CLDC [47]	Chinese	Joy, anger, surprise,fear,neutral, sadness
DES [48]	Danish	Anger, joy, sadness, surprise, neutral
EMODB [49]	German	Anger, joy, sadness, fear, disgust, boredom, and neutral
eNTERFACE [50]	English	Anger, disgust, fear, joy, surprise, sadness
FAU-AIBO [51]	German	joyful, surprised, emphatic, helpless, irritated, angry, bored, reprimanding, neutral, other
FERMUS III [52]	German, English	Anger, disgust, joy, neutral, sadness, surprise
GEMEP [53]	French	Amusement, pride, joy, relief, interest, hot anger, panic fear, despair, irritation, sadness, admiration, tenderness, disgust, contempt, anxiety, surprise, pleasure and neutral

Table 2.2. Affective speech corpora (cont.).

Corpus	Language	States /Traits
KES [54]	Korean	Neutral, joy, sadness, anger
KISMET [55]	English	Approval, attention, prohibition, soothing, and neutral
LDC EPSD [56]	English	Neutral, panic, anxiety, despair, sadness, hot anger, cold anger, elation, joy, interest, boredom, shame, pride, contempt
MPEG-4 [57]	English	Joy, anger, disgust, fear, sadness, surprise and neutral
Natural [58]	Mandarin	Anger, neutral
SAL [59]	English	Valence and activation
SmartKom [60]	German	Anger, helplessness, joy, neutral, pondering and surprise
SUSAS [61]	English	High stress, medium stress, neutral, scream
TESD [62]	Turkish	joy, surprise, sadness, anger, fear, neutral and other
Vera-Am-Mittag [63]	German	Valence, dominance and activation

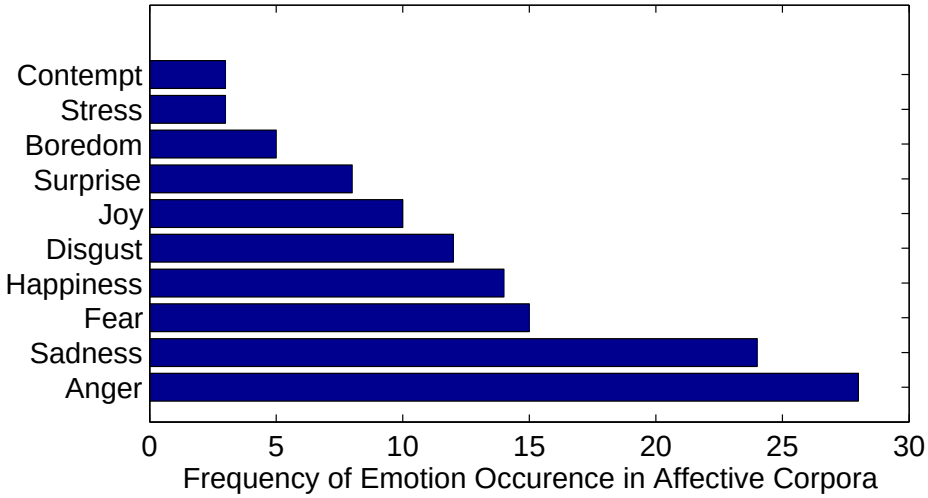

Figure 2.2. Frequency of emotion occurrence in affective speech corpora.

2.1.5. Review of Paralinguistic Challenges

The paralinguistic challenges have been instrumental in improving the state-of-the-art in this field. The challenges provide a great opportunity for the researchers in the field by bringing together experts from different disciplines, such as signal processing and psychology. Here we give a brief summary of the latest ComParE challenges. In INTERSPEECH 2012 the focus was speaker trait classification (personality, likeability and pathology) [64]. In 2013, ComParE challenge introduced four sub-challenges: Autism diagnosis, detection of social signals (laughter and fillers), prediction of level of conflict and classification of emotion [4], ComParE 2014 presented two sub-challenges, namely the prediction of physical load (labeled as high pulse/low pulse) and the ternary level of cognitive load [6]. The challenge organizers provided a baseline feature set (see Section 2.1.2), while the participants were allowed to use their own features and learning algorithm. The official performance measure in classification tasks is Unweighted Average Recall (UAR) in order to mitigate the class-imbalance problem. Firstly introduced in [65] as the competition measure, UAR can be defined as

$$UAR = \frac{1}{K}\sum_{k=1}^{K} TP(k)/P(k), \tag{2.1}$$

where K is the number of classes; $TP(k)$ and $P(k)$ denote the number of true positive instances and total positive instances for class k, respectively. The performance measure for the regression based challenges is either Mean Squared Error or correlation.

2.1.5.1. INTERSPEECH 2012 Speaker Trait Challenge. In 2012 Challenge, a baseline feature set, which is a subset of the one described in Section 2.1.2, was given to participants. Of the three challenges, the Likability Challenge is on prediction of how likable a subject is, and the Pathology Challenge measures intelligibility of speakers with laryngeal cancer on a read text. These two challenges had no sub-tasks. The personality challenge comprised prediction of five binary classification tasks one for each of Big-Five personality traits [66]: Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism (OCEAN), respectively. The contributions from 18 teams were accepted to the challenge. Six teams participated to all three challenges whereas 10 teams competed in only one of them. A detailed evaluation of the first challenge and survey on speaker traits can be found in [67]. Here, we briefly review top performing works in comparison to the baselines.

The organizers contrasted Support Vector Machines (SVM) and Random Forests (RF) for the baseline system, and reached test set UAR performances of 68.3% (average), 59.0% and 68.9% for personality, likeability and pathology challenges, respectively.

The winners of the personality challenge, Ivanov and Chen [68], proposed a system by augmenting the baseline set with modulation spectrum analysis features followed by Kolmogorov-Smirnov test based feature selection. The tailored features are trained with Adaboost meta classifier, reaching a test set average UAR performance of 69.3%.

The first runner up in the Personality Trait Challenge, Montacie and Caraty [69], combined three sub-systems composed of backward selected features from baseline set and proposed (pitch and intonation based) features in their best performing system that reached an averaged test set UAR of 68.4%, which is slightly over the baseline.

The test set performance of other systems did not outperform the challenge baseline.

The second runner up of this sub-challenge, Anumanchipalli *et al.* [70], fused four acoustic sub-systems based on frame-level Perceptual Linear Prediction (PLP) features (trained with a Multilayer Perceptron-MLP), utterance level baseline features (trained with SVM and RF) and frame-level baseline features (trained with SVM) in their system that obtained 68.1% UAR on the test set.

In the Likeability Challenge, work of Montacie and Caraty [69] outperformed the baseline reaching a top performance of 65.8% UAR. Adhering to the challenge protocol but not participating in the challenge, Brueckner and Schuller [71] used baseline openSMILE features training a set of Neural Network architectures. The highest performance (64.0% UAR on the test set) is obtained using a two-layer Neural Network with Gaussian-Bernoulli restricted Boltzmann machine at first stage. Buisman and Postma [72] proposed Log-Gabor based feature extraction, a common method in computer-vision, treating spectrograms as images. The Gabor-based LLD extraction is followed by descriptor level normalization, PCA reduction and RBF Kernel SVM training with gender dependent modeling, giving a test set UAR of 62.5%.

The highest improvement over the baseline performance is observed in the Pathology Sub-Challenge, where Kim *et al.* [73] attained 76.8% UAR by joint classification (by means of acoustic feature clustering and majority voting with each cluster) with feature combination of prosodic, intonational and five forward-selected features from baseline set. Lu and Sha [74] used Gaussian Processes for classification and regression, combining 13 kernels obtained from LLD based partitioning of baseline feature set. Kernel PCA is used for dimensionality reduction. This system reached a test set UAR performance of 73.7%. The second runner up of this sub-challenge, Huang *et al.* [75], reached their best result with fusion of three acoustic systems that are obtained with partitioning the baseline set into three sub-sets and applying asymmetric sparse partial least squares regression based feature reduction and training with LDA classifier. Table 2.3 summarizes the best UAR performances obtained by the top ranking studies in comparison with the baseline.

Table 2.3. Summary of top-performing works in the INTERSPEECH 2012 Speaker Trait challenge.

Rank	Work	UAR (%)
Personality Sub-Challenge		
	Baseline [64]	68.3
1	Ivanov and Chen [68]	69.3
2	Montacie and Caraty [69]	68.4
3	Anumanchipalli *et al.* [70]	68.1
Likeability Sub-Challenge		
	Baseline [64]	59.0
1	Montacie and Caraty [69]	65.8
2	Brueckner and Schuller [71]	64.0
3	Buisman and Postma [72]	62.5
Pathology Sub-Challenge		
	Baseline [64]	68.9
1	Kim *et al.* [73]	76.8
2	Lu and Sha [74]	73.7
3	Huang *et al.* [75]	71.9

2.1.5.2. INTERSPEECH 2013 Challenge. The INTERSPEECH 2013 Paralinguistic Challenge (ComParE 2013) consisted of four sub-challenges: Autism, Emotion, Conflict and Social Signal Sub-Challenge (SSC). Of the four, SSC was about frame-level detection (of laughter and fillers), and the remaining were utterance level-classification tasks. The challenge measure of SSC was Unweighted Average Area Under (Receiver Operating Characteristics) Curve (UAAUC). For the classification task based sub-challenges, UAR measure is used as in previous challenges. Papers from 13 teams were accepted to the challenge session, where the highest participation (7 papers) was for the Autism Sub-challenge. The baseline for the utterance level tasks is obtained using Linear SVM with 6 373 dimensional openSMILE feature set detailed in earlier sections. For SSC, a set of 47 popularly used Low Level Descriptors (MFCC 1-12, logarithmic energy, voicing probability, HNR, $F0$, and zero crossing rate) and their first order delta regression coefficients are extracted. For each frame, these LLDs are concatenated with mean and standard deviation of the features from the frame itself and its eight neighbors. A total of 141 features were provided as baseline for this task. The list of the top performing works in comparison with the baseline system is given in Table 2.4.

In the Autism Sub-Challenge, the top performance was obtained by Asgari *et al.* [76]. Their best test set performance, 69.4% UAR, was the only result outperforming the test set baseline of this sub-challenge (67.1%). This is partly attributed to the difficulty of the task and the competitiveness of the challenge baseline system. The authors used a system augmenting the baseline feature set with voice quality features extracted from proposed harmonic model. SVM classifier with the same setting as in the challenge baseline is used for model learning. In this challenge, Martinez *et al.* [77] extract their own prosodic features and process these features to obtain intermediary vectors (i-Vectors) and statistical descriptors. Diagnosis of autism spectrum disorder task is implemented with SVM classifier, reaching a UAR score of 66.1%. Lee *et al.* [78], combine machine learning algorithms such as SVM, deep neural networks (DNN) and weighted discrete k-nearest neighbors (WD-KNN) using baseline feature set. This work reaches a test set UAR of 64.8%, also remaining below the baseline.

For the Conflict Sub-Challenge, there were two papers accepted to the Chal-

lenge. Both of these works outperformed the baseline UAR of 80.8%. Rasanen and Pohjalainen [79] proposed Random Subset Feature Selection (RSFS) with K-Nearest Neighbor (KNN) classification, reaching 83.9% UAR performance. At each iteration, RSFS algorithm selects a random feature set and then measures relevance of each feature based on the performance of the subset that the feature participates in. To compute the relevance, the authors increase the weights of features participating in a set providing higher than average performance by a predefined value p, and similarly reduce the weight by the same amount for the features performing lower than the average. Grezes *et al.* [80] employ a system by first predicting speaker overlap from the baseline feature set and then stacking this feature to Linear SVM for conflict recognition. This two-level learning approach attains 83.1% UAR in the test set.

The Emotion Sub-challenge was yet another case where the baseline UAR of 40.9% on the 12-class classification task was highly competitive. The winner study of Goztzolya *et al.* [81] used two variants of Adaboost, namely AdaBoost.MH [82] and AdaBoost.MH:BA [83], via decision stumps/trees on the baseline feature set. This work attained a UAR of 42.3% on the challenge test set. Despite the laborious efforts, the best test score reached by the multiple learner combination study of Lee *et al.* [78] is 41.0% UAR, that gives an insignificant improvement over the baseline. Sethu *et al.* [84] investigate GMM and Joint Factor Analysis (JFA) based modeling to compensate speaker variability for acoustic emotion recognition. MFCCs (1-12) with their first order delta regression coefficients are used for modeling. Various alternative systems and their combinations are experimented giving over 48% UAR on the development set, however the best UAR performance reached on the test set is 35.7%.

The highest improvement over the baseline is observed in the SSC Sub-Challenge, where the baseline system does not fully take the benefit of time series property. Mainly benefiting from this property of the task, Gupta *et al.* [85] used smoothing and masking (suppressing the weak likelihood regions) on the probability outputs of a 2-hidden-layer DNN trained on baseline feature set. Out of the three parts, the highest contribution is attained from the use of DNN and smoothing, whereas the masking had a slight effect by reducing the false alarm rate. This work reported 91.5% UAAUC on the test

set. The first and second runners up have very similar UAR scores on the test set, 89.8% and 89.7% for Janicki [86] and Goztzolya *et al.* [81], respectively. Janicki [86] proposes fitting GMMs on three classes (laughter, filler and garbage) using MFCC based features, then applying SVM in the GMM log-likelihood space. The system proposed by Goztzolya *et al.* [81] is the same with their approach used in the Emotion Sub-challenge.

Table 2.4. Summary of top-performing works in the INTERSPEECH 2013 Challenge.

Rank	Work	Performance
Autism Sub-Challenge		
	Baseline [4]	67.1
1	Asgari *et al.* [76]	69.4
2	Martinez *et al.* [77]	66.1
3	Lee *et al.* [78]	64.8
Conflict Sub-Challenge		
	Baseline [4]	80.8
1	Rasanen and Pohjalainen [79]	83.9
2	Grezes *et al.* [80]	83.1
Emotion Sub-Challenge		
	Baseline [4]	40.9
1	Goztzolya *et al.* [81]	42.3
2	Lee *et al.* [78]	41.0
3	Sethu *et al.* [84]	35.7
Social Signals Sub-Challenge		
	Baseline [4]	83.3
1	Gupta *et al.* [85]	91.5
2	Janicki [86]	89.8
3	Goztzolya *et al.* [81]	89.7

2.1.5.3. INTERSPEECH 2014 Cognitive and Physical Load Challenge. The INTERSPEECH 2014 ComParE challenge [6] introduced two other paralinguistic aspects, namely cognitive and physical load, which are very important for tele-monitoring fa-

tigue of working people. Cognitive load is related to the strain put on the working memory, while the physical load is measured via heart pulse. The challenge presented three levels (low, medium and high) of cognitive load and two levels of physical load for classification. In vein with the previous challenges, UAR is used as performance measure.

The baseline system used is the same as INTERSPEECH 2013: 6 373 dimensional openSMILE features trained with Linear Kernel SVM. The baseline UAR scores of 71.9% and 61.6% were obtained on the test set for the Physical and Cognitive Load Sub-challenges (PLS/CLS), respectively.

Top performance on PLS (75.4% UAR) was obtained by Kaya *et al.* [10] using a multi-view discriminative projection based feature selection approach. Here, views correspond to LLD based feature groups. In each view, a feature ranking is obtained from the discriminative projection weights, and top ranking features from each view are combined for classification. Some views are pruned after ranking via Minimum Redundancy Maximum Relevance CCA Filter [9].

Van Segbroek *et al.* [87] analyze prosody (e. g. $F0$, silence ratio, intensity), separately model phoneme and word durations, and trial four different acoustic feature sets for i-Vector modeling. The study also benefits from speaker normalization, applied after speaker clustering. The fusion of four i-Vector systems with prosodic and phoneme statistics gives test set UAR scores of 68.9% and 73.9% for CLS and PLS, repsectively. This work ranks the first in CLS and the second in PLS.

Gostzolya *et al.* [88] employ two AdaBoost variants (as in their former study [81]), and Deep Rectifier Neural Networks (RELU-DNN) that learn faster and argued to generalize better without pre-training compared to sigmoid hidden unit based deep networks. The authors use an up-sampling strategy to overcome the class-imbalance and implement decision fusion (majority voting and posterior averaging) of RELU-DNN models to boost the accuracy. The study achieves 73.0% and 63.1% UAR for Physical and Cognitive Load Sub-challenges, respectively ranking third in both sub-challenges.

Focusing on CLS, in the work of Kua *et al.* [89] a set of GMM supervector based systems and their combinations are trialled with SVM, in addition to a frame-level baseline GMM system for classification. MFCC based features and Spectral Centroid Frequency [90] features are used as front-end. Alternative variability compensation systems including i-Vector and JFA modeling are used. Their best system which combines the baseline and two proposed feature types attain 63.7% UAR performance on the test set.

There are two other works that share the third rank in CLS apart from Gostzolya *et al.* [88], all reaching 63.1% UAR on the challenge test set. Huckvale [91] used Classification and Regression Trees (CART) and SVM on baseline plus VOQAL toolbox features with alternative methods for feature selection and normalization. This study highlights the importance of speaker dependent normalization, with excellent results on the development set using ground truth speaker meta-data. However, K-Means speaker clustering employed does not yield accurate results on the test set. Montacie and Caraty [92] proposed an SVM based classification of high-level speech features (e.g. speaking rate, pause ratio, filled pause ratio) summarized over the utterance using mean, standard deviation, kurtosis and skewness functionals, as well as the count of color words.

Evaluating the top contributions summarized in Table 2.5, we observe the importance of speaker normalization (or equivalently variability compensation), as well as the use of limited, but informative high level linguistic information (e. g. phoneme rate, duration of filled pauses). Fully unsupervised methods for speaker and linguistic content normalization/compensation need to be developed for robust real-life applications.

2.1.6. Open Issues

The field has grown fast in the last decade, however still some challenging/open issues remain, such as:

Table 2.5. Summary of top-performing works in the INTERSPEECH 2014 Challenge.

Rank	Work	UAR (%)
Pysical-Load Sub-Challenge		
	Baseline [6]	71.9
1	Kaya *et al.* [10]	75.4
2	Van Segbroek *et al.* [87]	73.9
3	Gostzolya *et al.* [88]	73.0
Cognitive Load Sub-Challenge		
	Baseline [6]	61.6
1	Van Segbroek *et al.* [87]	68.9
2	Kua *et al.* [89]	63.7
3	Gostzolya *et al.* [88] Huckvale [91] Montacie and Caraty [92]	63.1

- Collection of rich, annotated natural data
- Unsupervised, semi-supervised and cooperative learning methods to make use of large scale data with minimal annotation
- Cross-corpus, cross-language, and cross-domain affect recognition
- Real-time issues (robustness and applicability)
- Finding a compact set of descriptive/predictive features per paralinguistic task

Although there are notable acoustic emotion recognition studies on cross-corpus [3] and semi-supervised/cooperative learning [93], more studies are required in this direction. We proceed with the proposed feature filters targeting the last item.

2.2. Employed/Adapted Methods from Literature

2.2.1. Statistical Methods and Learners

2.2.1.1. Canonical Correlation Analysis. Proposed early in 1936 by Hotelling [94], CCA seeks to maximize the mutual correlation between two sets of variables by finding

linear projections for each set. Mathematically, CCA seeks to maximize the mutual correlation between two views of the same semantic phenomenon (e. g. audio and video of a speech) denoted $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times p}$, where n denotes the number of paired samples, via:

$$\rho(\boldsymbol{X}, \boldsymbol{Y}) = \sup_{\mathbf{w},\mathbf{v}} corr(\boldsymbol{X}\mathbf{w}, \boldsymbol{Y}\mathbf{v}), \tag{2.2}$$

where "corr" corresponds to Pearson's correlation, $\mathbf{w}$ and $\mathbf{v}$ correspond to the projection vectors of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Let $\mathbf{C_{XY}}$ denote the cross-set covariance between the sets X and Y, and similarly let $\mathbf{C_{XX}}$ denote within set covariance for X. The problem given in Equation 2.2 can be re-formulated as:

$$\rho(\boldsymbol{X}, \boldsymbol{Y}) = \sup_{\mathbf{w},\mathbf{v}} \frac{\mathbf{w}^T \mathbf{C_{XY}} \mathbf{v}}{\sqrt{\mathbf{w}^T \mathbf{C_{XX}} \mathbf{w} \cdot \mathbf{v}^T \mathbf{C_{YY}} \mathbf{v}}}. \tag{2.3}$$

The formulation in Equation 2.3 can be converted into a generalized eigenproblem for both projections (i. e. $\mathbf{w}$ and $\mathbf{v}$), the solution can be shown [95] to have the form of:

$$\mathbf{C_{XX}}^{-1} \mathbf{C_{XY}} \mathbf{C_{YY}}^{-1} \mathbf{C_{YX}} \mathbf{w} = \lambda \mathbf{w}, \tag{2.4}$$

where the correlation appears to be the square root of eigenvalue:

$$\rho(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{\lambda}. \tag{2.5}$$

To attain maximal correlation, the eigenvector corresponding to the largest eigenvalue in Equation 2.4 should be selected. Similarly, by restricting the new vectors to be uncorrelated with the previous ones, it can be shown that the projection matrices for each set are spanned by the k eigenvectors corresponding to the k largest eigenvalues. In short, when CCA is applied between $\boldsymbol{X}$ and $\boldsymbol{Y}$ we get:

$$[\boldsymbol{W}, \boldsymbol{V}, \rho, \boldsymbol{U}^X, \boldsymbol{U}^Y] = CCA(\boldsymbol{X}, \boldsymbol{Y}), \tag{2.6}$$

where $\boldsymbol{W}$ and $\boldsymbol{V}$ are composed of (sorted) eigenvectors from the eigenproblem in Equation 2.4, r is the m dimensional vector of canonical correlations given in Equation 2.5 while $\boldsymbol{U}^X$ and $\boldsymbol{U}^Y$ are the covariates. In other words, $\boldsymbol{U}^X = \boldsymbol{X} \times \boldsymbol{W}$, when features in $\boldsymbol{X}$ are mean removed. The relationship between the canonical correlation and the corresponding covariates is given by the the Pearson's Correlation Coefficient (PCC):

$$\rho_i = PCC(\boldsymbol{U}_i^X, \boldsymbol{U}_i^Y), \tag{2.7}$$

where i indexes the columns. It is important to note that the maximum number of covariates m in $\boldsymbol{U}^X$ and $\boldsymbol{U}^Y$ are limited with the matrix rank of $\boldsymbol{X}$ and $\boldsymbol{Y}$:

$$m = min(rank(\boldsymbol{X}), rank(\boldsymbol{Y})) \tag{2.8}$$

Non-linearity can be incorporated into CCA using the *kernel trick* [95] or deep neural networks [96]. CCA is related to two extensively used statistical methods: PCA and LDA [97]. When CCA is applied between feature matrix and the identity matrix, the result is PCA. Bartlett showed that LDA is a special case of CCA, and can be obtained when CCA is applied between feature matrix and 1-of-K coded label matrix [98]. Another recent method called Covariance Operator Inverse Regression [99] is proved to give identical central space with Kernel CCA.

2.2.1.2. CCA for Regression. CCA is commonly used for several tasks ranging from covariate extraction, to modality/representation fusion [100]. It has also applications in ranking feature selection [9, 101]. However, although it can be used for non/linear regression, this use is not common to the best of our knowledge. Among few studies in this vein, Nicolaou *et al.* introduce Correlated-Spaces Regression (CSR) inspired from CCA and the high inter-correlation of emotion dimensions [102].

To employ CCA as a regressor using Equation 2.2 , we note that as the correlation

is 1, the covariates of the two views $\boldsymbol{X}$ and $\boldsymbol{Y}$ become identical:

$$\rho(\boldsymbol{X}, \boldsymbol{Y}) = \sup_{\mathbf{w},\mathbf{v}} corr(\boldsymbol{X}\mathbf{w}, \boldsymbol{Y}\mathbf{v}) = 1 \iff \boldsymbol{X}\mathbf{w} = \boldsymbol{Y}\mathbf{v}, \tag{2.9}$$

where we assume both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are mean-normalized. When correlation is close to 1, we can use one representation to reconstruct the other. Let column vectors $\mu_{\mathbf{X}}$ and $\mu_{\mathbf{Y}}$ denote respective training set means used to normalize the sets, Equation 2.9 can be rewritten as

$$(\boldsymbol{Y} - \mathbf{1}\mu_{\mathbf{Y}}^{\mathbf{T}})\mathbf{v} \simeq (\boldsymbol{X} - \mathbf{1}\mu_{\mathbf{X}}^{\mathbf{T}})\mathbf{w}, \tag{2.10}$$

where $\mathbf{1}$ is a column vector of ones having length equal to respective dataset cardinality. It is straightforward to convert Equation 2.10 to reconstruct one side (here Y):

$$\boldsymbol{Y} \simeq (\boldsymbol{X} - \mathbf{1}\mu_{\mathbf{X}}^{\mathbf{T}})\mathbf{w}\mathbf{v}^{\dagger} + \mathbf{1}\mu_{\mathbf{Y}}^{\mathbf{T}}, \tag{2.11}$$

where † denotes a generalized inverse. Note that in the case of regression v is a scalar. This approach can be used in regression where one view represents the target variables.

2.2.1.3. Local Fisher Discriminant Analysis. It is known that when classes are multimodal, FDA faces anomalies [103]. It is important to preserve the local structure in the embedded space while trying to maximize the class separability. To retain the multimodality in the target space without regarding the classes, Locality Preserving Projection (LPP) [104] is introduced as an alternative to Principal component Analysis. The approach uses the affinity matrix idea to weight (softly mask) the projections. This idea inspired Sugiyama to extend traditional FDA to Local FDA by first reformulating the scatter matrices [105]:

$$S^{w} = 1/2 \sum_{i,j}^{n} A_{i,j}^{w}(x_i - x_j)(x_i - x_j)', \tag{2.12}$$

$$S^{b} = 1/2 \sum_{i,j}^{n} A_{i,j}^{b}(x_i - x_j)(x_i - x_j)', \tag{2.13}$$

where ($'$) denotes transpose and

$$A^w_{i,j} = \begin{cases} 1/n_c & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{2.14}$$

$$A^b_{i,j} = \begin{cases} 1/n - 1/n_c & \text{if } y_i = y_j = c, \\ 1/n & \text{if } y_i \neq y_j, \end{cases} \tag{2.15}$$

Here the affinity matrices do not contain locality information but class information. To obtain LFDA we have [105]:

$$\overline{\mathcal{S}}^w = 1/2 \sum_{i,j}^{n} \overline{\mathcal{A}}^w_{i,j} (x_i - x_j)(x_i - x_j)', \tag{2.16}$$

$$\overline{\mathcal{S}}^b = 1/2 \sum_{i,j}^{n} \overline{\mathcal{A}}^b_{i,j} (x_i - x_j)(x_i - x_j)', \tag{2.17}$$

and localized discriminative affinity matrices are defined as

$$\overline{\mathcal{A}}^w_{i,j} = \begin{cases} A_{i,j}/n_c & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \tag{2.18}$$

$$\overline{\mathcal{A}}^b_{i,j} = \begin{cases} A_{i,j}(1/n - 1/n_c) & \text{if } y_i = y_j = c, \\ 1/n & \text{if } y_i \neq y_j, \end{cases} \tag{2.19}$$

where $A_{i,j}$ is the $n \times n$ regular affinity matrix keeping the unsupervised locality information. $A_{i,j}$ can simply be composed of 1s for k-nearest neighbors for each instance and 0s for the rest. It is also possible to adopt a localized measure where the distance to the k-th nearest neighbor is used as bandwidth in Gaussian similarity. Let D denote the $n \times n$ Euclidean distance matrix of samples, d_k is the n dimensional vector keeping the square root of the Euclidean distance of each sample to its k-th neighbor, and $M./L$ denote the element-wise division, we can obtain a smoother affinity matrix A via:

$$L = d_k d_k', \tag{2.20}$$

$$A = \exp(-D./L). \tag{2.21}$$

Once the scatter matrices are computed, the regular FDA eigenproblem can be used to obtain the discriminative projection:

$$\overline{\mathcal{S}}^{b} W = \Lambda \overline{\mathcal{S}}^{w} W. \tag{2.22}$$

2.2.1.4. Extreme Learning Machines. The Extreme Learning Machine (ELM) classifier was first introduced in [106] as a fast alternative training method for Single Layer Feedforward Networks (SLFNs). The rigorous theory of the ELM paradigm is presented in 2006 by Huang *et al.* [13], where the authors compare the performance of ELM, SVM, and Back Propagation (BP) learning based SLFN in terms of training time and accuracy. The basic ELM paradigm has matured over the years to provide a unified framework for regression and classification, related to generalized SLFN class including Least Square SVM (LSSVM) [14, 107].

Despite the speed and accuracy of ELMs, they were only recently employed in affective computing exhibiting outstanding performance with typically under-sampled high dimensional datasets [17, 108]. In one of the recent studies, Han *et al.* [108] use DNNs for extraction of higher level features (class distribution) from segment based acoustic descriptors, then summarize these features over the utterances using simple statistical functionals (e. g. mean, max). The suprasegmental features were stacked as input to ELMs. They show that ELM based systems outperform both SVM and HMM based systems.

The argument of the basic ELM introduced by Huang *et al.* is that the first layer (input layer) weights and biases of a neural network classifier do not depend on data and can be randomly generated, whereas the second layer (output weights) can be effectively and efficiently solved via least squares [13]. The input layer can be considered as carrying out unsupervised feature mapping, and the activation function outputs (the output matrix) are subjected to a supervised learning procedure. Let $(\mathbf{W}, \mathbf{b}, \mathbf{H}, \beta)$ denote an SLFN, where the output with respect to input $\boldsymbol{X} \in \mathbb{R}^d$ is given as $\hat{\boldsymbol{Y}} = \mathbf{h}(\mathbf{W}\mathbf{x} + \mathbf{b})\beta$. Here, $\mathbf{W}$ and $\mathbf{b}$ denote the randomly generated mapping

matrix, and the bias vector, respectively. $\mathbf{h}(\mathbf{x}) \in \mathbb{R}^p$ denotes the hidden node output and $\mathbf{H} \in \mathbb{R}^{N \times p}$ denotes the hidden node output matrix. β is the analytically learned second layer weight matrix.

The nonlinear activation function $h()$ can be any infinitely differentiable bounded function. A common choice for $h()$ is the sigmoid function:

$$h(\mathbf{a}) = \frac{1}{1 + exp\,(-\mathbf{a})}. \tag{2.23}$$

ELM proposes unsupervised, even random generation of the hidden node output matrix $\mathbf{H}$. The actual learning takes place in the second layer between $\mathbf{H}$ and the label matrix $\mathbf{T}$. $\mathbf{T}$ is composed of continuous annotations in case of regression, therefore is a vector. In the case of K-class classification, $\mathbf{T}$ is represented in one vs. all coding:

$$\mathbf{T}_{t,k} = \begin{cases} +1 & \text{if } y^t = k, \\ -1 & \text{if } y^t \neq k. \end{cases} \tag{2.24}$$

The second level weights β are learned by least squares solution to a set of linear equations $\mathbf{H}\beta = \mathbf{T}$. Proving first that random projections and nonlinear mapping with $L \leq N$ result in a full rank $\mathbf{H}$, the output weights can be learned via

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \tag{2.25}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse [109] that gives the minimum L_2 norm solution to $||\mathbf{H}\beta - \mathbf{T}||$, simultaneously minimizing the norm of $||\beta||$. It is important to mention that ELM is related to Least Square SVMs via the following output weight learning formulation:

$$\beta = \mathbf{H}^T(\frac{\mathbf{I}}{\tau} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{T}, \tag{2.26}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix, and τ, which is used to regularize the linear

kernel $\mathbf{H}\mathbf{H}^T$, is indeed the complexity parameter of LSSVM [107]. The approach is extended to use any valid kernel. A popular choice for kernel function is Gaussian (RBF):

$$K(\mathbf{x_k}, \mathbf{x_l}) = \phi(\mathbf{x_k}) \cdot \phi(\mathbf{x_l}) = \exp(-\frac{||\mathbf{x_k} - \mathbf{x_l}||}{\sigma^2}) \tag{2.27}$$

In both (basic and kernel) approaches, the prediction of $\boldsymbol{X}$ is given via $\hat{\boldsymbol{Y}} = \mathbf{h}(\mathbf{x})\beta$. In case of multi-class classification, the class with maximum score in $\hat{\boldsymbol{Y}}$ is selected. Inspired by the success of SLFN based auto-encoders for feature enhancement and the relationship between Principal Component Analysis (PCA) and SLFNs [28], in this thesis, we further consider the use of PCA instead of random generation of input weights.

2.2.1.5. Partial Least Squares. PLS regression between two sets of variables $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ and $\boldsymbol{Y} \in \mathbb{R}^{N \times p}$ is based on decomposing the matrices as $\boldsymbol{X} = \mathbf{U}_x\mathbf{V}_x + r_x$, $\boldsymbol{Y} = \mathbf{U}_y\mathbf{V}_y + r_y$, where $\mathbf{U}$ denotes the latent factors, $\mathbf{V}$ denotes the loadings and r stands for the residuals. The decomposition is done by finding projection weights $\mathbf{W}_x, \mathbf{W}_y$ that jointly maximize the covariance of corresponding columns of $\mathbf{U}_x = \boldsymbol{X}\mathbf{W}_x$ and $\mathbf{U}_y = \boldsymbol{Y}\mathbf{W}_y$. Further details of PLS regression can be found in [12]. PLS is applied to classification in one-versus-all setting between the feature matrix $\boldsymbol{X}$ and the binary label vector $\boldsymbol{Y}$, then the class giving the highest regression score is taken as prediction. The number of latent factors is a hyper-parameter to tune via cross-validation.

2.2.2. Visual Feature Extraction

2.2.2.1. Local Phase Quantization. The LPQ features are computed by taking 2-D Discrete Fourier Transform (DFT) of M-by-M neighborhood of each pixel in the gray scale image. In our implementation we use $M = 9$. 2D-DFT is computed at four frequencies $\{[a, 0]^T, [0, a]^T, [a, a]^T, [a, -a]^T\}$ with $a = 1/M$, which correspond to four of eight neighboring frequency bins centered at the pixel of interest. The real and

imaginary parts of resulting four complex numbers are separately quantized using a threshold of zero, that gives an eight bit string. The eight bit string is then converted into an integer value in the range 0-255. The pixel based values are finally converted into a histogram of 256 bins. Since this histogram representation does not keep the structural information of facial features, the face is divided into non-overlapping regions and an LPQ histogram is computed per region.

2.2.2.2. Local Binary Patterns from Three Orthogonal Planes. After face alignment and conversion to gray scale, LBP computation amounts to finding the sign of difference with respect to a central pixel in a neighborhood, transforming the binary pattern into an integer and finally converting the patterns into a histogram. Uniform LBP clusters 256 patterns into 59 bins, and takes into account occurrence statistics of common patterns [110]. As in LPQ, the face is divided into non overlapping regions and an LBP histogram is computed per region. The TOP extension applies the relevant descriptor on XY, XT and YT planes (T represents time) independently and concatenates the resulting histograms.

2.2.2.3. Local Gabor Binary Patterns from Three Orthogonal Planes. In LGBP-TOP, the images are convolved with a set of 2D complex Gabor filters to obtain Gabor-videos, then LBP-TOP is applied to image blocks from each Gabor-video. A 2D complex Gabor filter is the convolution of a 2D sinusoid (carrier) having phase P, spatial frequencies u_0 and v_0 with a 2D Gaussian kernel (envelope) having amplitude K, orientation θ,and spatial scales a and b. In line with [111], for simplicity we take $a = b = \sigma, u_0 = v_0 = \phi$ and $K = 1$ to obtain

$$G(x, y) \exp(-\pi\sigma^2((x - x_0)_r^2 + (y - y_0)_r^2)) \exp(j(2\pi\phi(x + y) + P)), \tag{2.28}$$

Figure 2.3. Illustration of an aligned image and two of its Gabor magnitude response pictures.

where the subscript $_r$ stands for a clockwise rotation operation around reference point (x_0, y_0) such that

$$\begin{aligned}(x - x_0)_r &= (x - x_0)cos\theta + (y - y_0)sin\theta \\ (y - y_0)_r &= -(x - x_0)sin\theta + (y - y_0)cos\theta\end{aligned} \tag{2.29}$$

Note that the effect of the phase is canceled out, since only the magnitude response of the filter is used for the descriptor. A sample video image with Gabor magnitude response images are given in Figure 2.3.

When 2D complex Gabor filters are formed, all video frames are convolved and separate Gabor-videos are stacked to LBP-TOP operation. For LBP-TOP computation, we use non-overlapping blocks of 4 frames and divide all planes (i. e., XY, XT and YT) into 16 non-overlapping, equal-size regions. Also in our implementation, we divide the video into two equal length volumes over the time axis and extract LGBP-TOP features from each volume to further enhance temporal modeling.

2.2.2.4. Video Modeling in Riemannian Manifold. The image features are represented in a variety of ways over the video. Here, three alternative approaches are given for video (image set) modeling. All of these representations are known to lie in Riemannian manifolds, therefore regular vector space operations can not be applied. Thus,

appropriate kernels are used to measure similarity of videos.

The first approach is taking Singular Value Decomposition (SVD) of the video feature matrix X. Let r be the rank of matrix X. SVD gives an orthonormal decomposition in the form

$$X = U\Lambda V^T, \tag{2.30}$$

where columns of U are normalized eigenvectors of XX^T, rows of V^T are normalized eigenvectors of X^TX, and first r diagonal elements of Λ are square the roots of corresponding sorted eigenvalues. Representing the video with the first $l \leq r$ columns of U leading to a matrix $L \in \mathbb{R}^{d\times l}$. This linear subspace representation is known to lie in Grassmanian manifold G(l,d), which is a special case of Riemannian manifold [112].

Our second approach represents the image set $X_v \in \mathbb{R}^{d\times F_v}$ with its $d\times d$ covariance matrix Σ. The third extends this by introducing the mean statistic μ of the features, thus obtaining a multivariate Gaussian. To embed the Gaussian in a Riemannian manifold, it is represented as a Symmetric Positive Definite (SPD) matrix [113]:

$$\mathcal{N}(\mu,\Sigma) \sim M = |\Sigma|^{-\frac{1}{d+1}} \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix} \tag{2.31}$$

Since regular vector space operations so not hold on these matrices, Riemannian kernels are used to compute video similarity. For the SVD based linear subspace, the similarity of the video matrices L_i and L_j is computed via Mercer kernels that map the points in a Grassmanian manifold to Euclidean space [112, 114]. Linear projection kernel $\mathcal{K}_{i,j}^{Proj.-Lin.}$ is defined as

$$\mathcal{K}_{i,j}^{Proj.-Lin.} = ||L_i^T L_j||_F^2, \tag{2.32}$$

where $||.||_F^2$ is the Frobenus norm. The RBF kernel is defined over the mapping $\Phi_{Proj.} =$

$L_i L_i^T$ [115]:

$$\mathcal{K}_{i,j}^{Proj.-RBF.} = \exp(-\gamma||L_i L_i^T - L_j L_j^T||_F^2) \tag{2.33}$$

The Covariance matrix and the Gaussian matrix representation of videos are both Symmetric Positive Definite (SPD). A popular distance measure for SPD matrices is the Log-Euclidean Distance (LED), which is based on a matrix logarithm operator [116]. The proposed Linear and RBF kernels between SPD matrices S_i and S_j are formulated as [115, 117]:

$$\mathcal{K}_{i,j}^{LED-Lin.} = trace\,[log(S_i)log(S_j)] \tag{2.34}$$

$$\mathcal{K}_{i,j}^{LED-RBF.} = \exp(-\gamma trace\,[log(S_i) - log(S_j)]) \tag{2.35}$$

While the obtained kernels can be given to kernel machines as input, they can also suitably be used in other learners, where similarity to training instances is considered as a new feature representation. In this study, we train models using Partial Least Squares (PLS) and ELM on the obtained kernels.

3. PROPOSED METHODS

In this chapter, we present our novel discriminative projection based feature filters and automatic mixture model selection algorithm in Sections 3.1 and 3.2, respectively. Then, in Section 3.3, we draw the links between these two branches using the relationship between the base statistical methods; and explain how the relationship can be exploited in future studies.

3.1. Discriminative Projection Based Feature Filters

Throughout the thesis, five different of discriminative projection based feature filters are proposed. We first introduced several CCA based feature selection (FS) alternatives, including minimum Redundancy Maximum Relevance Filter [9]. However, computationally the least costly approach called Samples versus Labels CCA Filter (SLCCA-Filter) was found to be the most successful. Then we developed a multi-view approach using LLD information to partition the massive feature set [10]. Here, discriminative projections are applied to each view separately and the top ranking k features from each view are concatenated for model learning. Lastly, a randomized version of SLCCA-Filter that does not necessitate domain knowledge is proposed [11]. Note that the proposed FS methods can also be used with other discriminative projections, such as Local Fisher Discriminant Analysis (LFDA) [105]. The reason we chose CCA is that it is flexible to be used with continuous (regression) and discrete (classification) target variables. It can also be employed in a semi-supervised or weakly supervised setting for feature reduction [118, 119]. We first provide background on CCA, then give the details of CCA based filter methods in the following subsections.

3.1.1. Samples versus Labels CCA Filter

The main idea behind the Samples versus Labels CCA Filter (SLCCA-Filter) algorithm we proposed in [9] is as follows. When features in one view are subjected to CCA against the labels on the other view, the absolute value of the projection matrix

$\boldsymbol{W}$ can be used to rank the features. The application to regression is straightforward, since the resulting matrix is $n \times 1$, therefore a vector. It can be applied in the same way to binary classification, where the classes can be denoted with 0 and 1 in the target. For $K > 2$, we can use the canonical correlation value (ρ_i) to weight the corresponding projection column (eigenvector $\boldsymbol{W}_i$). In short, the SLCCA-Filter algorithm, which takes as inputs a dataset $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and a label matrix $\boldsymbol{T} \in \{0,1\}^{n \times (K)}$ is given as:

$$[\boldsymbol{W}, \boldsymbol{V}, \rho] = CCA(\boldsymbol{X}, \boldsymbol{T}), \tag{3.1}$$

$$\boldsymbol{h} = \sum_{i=1}^{m} abs(\boldsymbol{W}_i)\rho_i, \tag{3.2}$$

$$\mathbf{R} = \text{sort}(\boldsymbol{h}, \text{`descend'}), \tag{3.3}$$

where the 1-of-K coded label matrix $\boldsymbol{T}$ is defined as

$$T_{j,k} = \begin{cases} 1 & \text{if } y_j = k, \\ 0 & \text{if } y_j \neq k, \end{cases} \tag{3.4}$$

and $\mathbf{R}$ is the output of feature ranking. Here, j and k index the instances and the classes, respectively. Since K classes have $K-1$ degrees of freedom, the rank of matrix $\boldsymbol{T}$ is $K-1$. Therefore it is possible to remove any of the columns from the 1-of-K coded matrix. The filter can be applied to Fisher Discriminant Analysis (FDA) or to its localized version LFDA [105] in a similar manner, as we have shown in our recent study [10]. In FDA variants, instead of ρ^i, the square root of the corresponding eigenvalue λ^i is used to weight the projection matrix. Note that the approach can also be used for multi-task learning, both in classification and regression.

3.1.2. Random Discriminative Projection Based Filters

Although it looks efficient in suppressing redundant features, SLCCA-Filter has an important drawback, which gives the motivation to the research described in this section: the number of non-zero weight features in the projection is upper-bounded by the rank of the data matrix. When $d > n$, a pseudo-inverse operation takes the place of

the inverse for the singular covariance matrix, and subsequently valuable information is lost.

By means of random sampling of features, it is possible to evaluate feature relevance/redundancy in different conditions and aggregate them to obtain a final ranking. While the absolute value of feature projection matrix provides information about feature level importance (driven to zero if the feature is redundant or irrelevant), the square root of the eigenvalue in a discriminative projection can be used to weight how good the feature group collectively performs.

In our approach, at each iteration we project a random subset of $d/2$ features and its complement set, where d is dataset dimensionality. We then aggregate the absolute value of projection weights multiplied with corresponding eigenvalues. After L iterations, the accumulated feature importance vector is sorted in descending order to provide the ranking. With this approach, we can both access all features at each iteration, and also obtain compatible feature weights in the projection matrix. The proposed algorithm is given in Figure 3.1.

If we only select $p << d$ for the projection without the complement set (as in the case of Random Forests [120]) the algorithm needs hundreds of iterations to include the majority of features. When p of d features are sampled, at iteration t, the ratio of unselected features (the probability that a feature is never selected), is $(1 - p/d)^t$. Figure 3.2 illustrates the expected ratio of unselected features with respect to the number of sampling iterations for an imaginary dataset of dimensionality $d = 10\,000$. From the figure, we see that with a lower range values for t and p, which are popularly used in RFs, a high proportion of features are never evaluated. Thus, this approach either leaves a markedly high amount of features that can be potentially beneficial unattended, or requires a high number of iterations to include them. In the first case the classification performance is compromised, whereas in the latter the computational load is higher.

Thus, the proposed algorithm is expected to give better performance with much

Input:

$\boldsymbol{X}$: n x d dimensional data matrix

$\boldsymbol{T}$: target matrix n x (K-1) in classification,

n x 1 in regression

L : Number of iterations

Require $\boldsymbol{X}$, $\boldsymbol{T}$ and L as input.

p $\Leftarrow$ d / 2;

$\boldsymbol{w} \Leftarrow$ zeros(d,1);

for $i = 1$ to L **do**

RandFeats $\Leftarrow$ randperm(d);

Feats $\Leftarrow$ RandFeats(1:p);

$\boldsymbol{X}_{rand} \Leftarrow \boldsymbol{X}$(:, Feats);

$\overline{\boldsymbol{X}_{rand}} \Leftarrow \boldsymbol{X}$(:, Complement Set of Feats);

Apply CCA between feature set $\boldsymbol{X}_{rand}$ and $\boldsymbol{T}$ using Equation 3.1;

Apply CCA between complement set $\overline{\boldsymbol{X}_{rand}}$ and $\boldsymbol{T}$ using Equation 3.1;

Compute the weight vectors $\boldsymbol{h}_i$ and $\overline{\boldsymbol{h}_i}$ for each projection using Equation 3.2;

Combine $\boldsymbol{h}_i$ and $\overline{\boldsymbol{h}_i}$ to obtain d dimensional importance vector $\boldsymbol{w}_i$;

Accumulate weighted features in vector $\boldsymbol{w}$: $\boldsymbol{w} \Leftarrow \boldsymbol{w} + \boldsymbol{w}_i$;

end for

Get sorted weights,$\boldsymbol{w}_s$ and feature ranking $\mathbf{R}$ by applying Equation 3.3 on $\boldsymbol{w}$;

Figure 3.1. Random SLCCA algorithm.

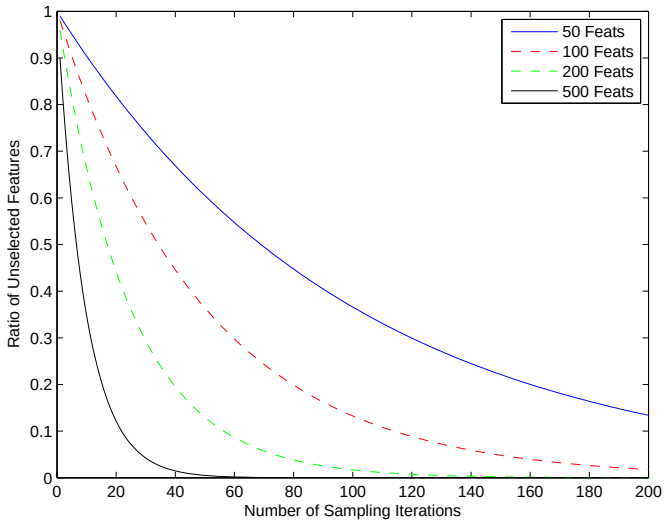

Figure 3.2. Illustration of the expected proportion of unselected features (the probability of not selecting a feature) after t iterations with varying number of sampled features (p).

fewer iterations compared to sampling in Random Forests fashion. In our preliminary experiments, we verified that selecting $p << d$ results in poor performance when small values for p and L (in the range [10-100]) are used. On the other hand, if there is a high discrepancy between the dimensionality of the random subset and its complement, the weights are incompatible, and the ranking would be misleading.

3.1.3. Minimum Redundancy Maximum Relevance CCA

We next propose the minimum Redundancy Maximum Relevance CCA (mRMR-CCA), which is related to CFS [121] and mRMR [122]. The difference from these methods is that instead of computing feature-wise internal correlations and averaging results, we directly compute canonical correlation of a candidate feature against the already selected subset:

$$\arg \max_{x_j \in X - S_{k-1}} \left[\rho_{CCA}(x_j, t) - \rho_{CCA}(x_j, S_{k-1}) \right], \tag{3.5}$$

where S_{k-1} denotes the already selected subset with $k-1$ features, x_j is a candidate feature and t is the target variable (label). In Equation 3.5 the subtraction operator can be replaced with division, to account for the relative merit with respect to internal correlations. In our experiments, we used subtraction based measures.

3.1.4. Maximum Collective Relevance CCA

Our last CCA based feature filter, Maximum Collective Relevance CCA (MCR-CCA) focuses on maximizing the joint correlation of the selected subset and the candidate feature against the target. The redundancy in the ranked subset can further be reduced using feature extraction. The formulation is similar to wrapper based forward selection [123], but we do not employ a classifier:

$$\arg \max_{x_j \in X - S_{k-1}} \left[\rho_{CCA}(S_{k-1} \cup x_j, t) \right]. \tag{3.6}$$

3.1.5. Relation to Previous Work

One of the novel methods we introduce, namely mRMR-CCA, is related to CFS [121] and mRMR [122]. CFS measures the heuristic merit between a feature set S and target t via [121]:

$$r_{S,t} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}, \tag{3.7}$$

where k is number of features, $\overline{r_{zi}}$ denote average correlation between the features in the subset and the target variable, and the term $\overline{r_{ii}}$ denote average inter-correlation between features. In short, Equation 3.7 punishes internal correlation and favors higher average feature-target correlations. Hall (1999) proposes several measures of dependence to compute feature-feature and feature-target merits of a subset. When the target variable is continuous, Pearson's correlation coefficient is used. In our approach we simplify Equation 3.7, keeping the notion of high relevance low redundancy. Similarly, mRMR

drives the feature selection (FS) in a set X, at step k maximizing the difference or ratio between relevance and redundancy terms [122]:

$$\arg \max_{x_j \in X - S_{k-1}} \left[I(x_j, t) - \frac{1}{k-1} \sum_{x_i \in S_{k-1}} I(x_j, x_i) \right], \tag{3.8}$$

where $I(x, y)$ is mutual information between random variables x and y. In KCCAm-RMR, Sakar *et al.* [101] improved mRMR FS using correlated functions of variables (i. e. projections attained by CCA) weighted with corresponding correlations with the target variable. In our work, we completely replace MI with CCA. Moreover, CCA not only eliminates discretization for continuous targets, but also is capable of handling multiple targets in the feature reduction process.

An FS method similar to SLCCA-Filter is previously used by Hardoon *et al.* [124] to determine correlated pixels in fMRI analysis.

The proposed FS methods are applied on paralinguistic challenge corpora giving state-of-the-art results in prediction of depression severity level (Section 5.3), physical load of the speaker (Section 4.3) and conflict level in dyadic interactions (Section 4.2).

3.2. Adaptive Mixture of Factor Analyzers

Mixture models have a widespread use in various domains of machine learning and signal processing for supervised, semi-supervised and unsupervised tasks [125,126]. However, the model selection problem remains to be one of the challenges and there is a need for efficient and parsimonious automatic model selection methods [127].

Let $\boldsymbol{x}$ denote a random variable in $\mathbb{R}^d$, a mixture model represents the distribution of $\boldsymbol{x}$ as a mixture of K component distributions:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p\left(\boldsymbol{x}|\mathcal{G}_k\right) p\left(\mathcal{G}_k\right), \tag{3.9}$$

where $\mathcal{G}_k$ correspond to components, and $p(\mathcal{G}_k)$ are the prior probabilities of the components. The likelihood term, expressed by $p(\boldsymbol{x}|\mathcal{G}_k)$, can be modeled by any distribution. The most commonly employed mixture model in pattern recognition and speech signal processing is the Gaussian Mixture Model:

$$p(\boldsymbol{x}|\mathcal{G}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.10}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean and covariance of the k^{th} component distribution, respectively. However, GMMs suffer the curse of dimensionality heavily, therefore are either used with a small set of features with full covariance or with only diagonal covariance, which neglects important feature correlations that are very useful in speech processing. It is possible to keep a low number of parameters for the model without sacrificing correlation information by adopting a factor analysis approach. Factor Analysis (FA) is a latent variable model, which assumes the observed variables are linear projections of a small number of independent factors $\boldsymbol{z}$ with additive Gaussian noise:

$$\boldsymbol{x} = \boldsymbol{\Lambda}\boldsymbol{z} + \boldsymbol{u}, \boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}), \boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{\Psi}), \tag{3.11}$$

where $\boldsymbol{\Lambda}$ is a $d \times p$ *factor loading matrix* and $\boldsymbol{\Psi}$ is a diagonal uniquenesses matrix representing the common sensor noise. Subsequently, the covariance matrix in Equation 3.10 is expressed as $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^T + \boldsymbol{\Psi}$, effectively reducing the number of parameters from $O(d^2)$ to $O(dp)$, with $p << d$. If each Gaussian component is expressed in the latent factor space, the result is a mixture of factor analyzers (MoFA). Both for FA and MoFA Expectation-Maximization (EM) algorithm [128] based methods exist to infer model parameters [126, 129], but these approaches require the specification of hyper-parameters like the number of clusters and the number of factors per component. For the model selection problem of MoFA, an incremental algorithm (IMoFA) was proposed in [130], where factors and components were added to the mixture one by one. The model complexity was monitored on a separate validation set.

In this section, we propose a fast and parsimonious model selection algorithm

called *Adaptive Mixture of Factor Analyzers* (AMoFA). Similar to IMoFA, AMoFA is capable of adapting a mixture model to data by selecting an appropriate number of components and factors per component. In particular, the proposed AMoFA algorithm deals with two shortcomings of the IMoFA approach: (i) Instead of relying on a validation set, AMoFA uses a Minimum Message Length (MML) based criterion to control model complexity, subsequently using more training samples in practice. (ii) AMoFA is capable of removing factors and components from the mixture when necessary.

The remainder of this section is organized as follows. In the next section we briefly summarize the related literature work and methods before detailing AMoFA in Section 3.2.2. In Section 3.2.3 we illustrate the automatic model selection capability on synthetic low dimensional data. Comparative experiments on high dimensional classification problems via class conditional modeling are given in Section 3.2.4. Section 3.2.5 presents an application of AMoFA to acoustic child emotion recognition, whereas Section 3.2.6 gives an overview with future directions.

3.2.1. Related Work

There are numerous studies for mixture model class selection. These include using information theoretical trade-offs between likelihood and model complexity [131–135], greedy approaches [130, 136] and full Bayesian treatment of the problem [137–140]. A brief review of related automatic model selection methods is given in Table 3.1, a detailed treatment can be found in [141].

In one of the most popular model selection approaches for Gaussian mixture models (GMMs), Figueiredo and Jain proposed to use an MML criterion for determining the number of components in the mixture, and shown that their approach is equivalent to assuming Dirichlet priors for mixture proportions [143]. In their method, a large number of components (typically 25-30) is fit to the training set, and these components are eliminated one by one. At each iteration, the EM algorithm is used to find a converged set of model parameters. The algorithm generates and stores all intermediate models, and selects one that optimizes the MML criterion.

Table 3.1. Automatic mixture model selection approaches.

Work	Model Selection	Approach
Pelleg & Moore (2000) [142]	MDL	Incremental
Ghahramani & Beal (2000) [137]	Variational Bayes	Incremental
Rasmussen (2000) [138]	MC for DPMM	Both
Figueiredo & Jain (2002) [143]	MML	Decremental
Verbeek *et al.* (2003) [136]	Fixed iteration	Incremental
Law *et al.* (2004) [144]	MML	Decremental
Zivkovic & v.d. Heijden (2004) [145]	MML	Decremental
Salah & Alpaydin (2004) [130]	Cross Validation	Incremental
Constantinopoulos *et al.* (2007) [146]	Variational Bayes	Incremental
Gomes *et al.* (2008) [139]	Variational DP	Incremental
Boutemedjet *et al.* (2009) [147]	MML	Decremental
Gorur & Rasmussen (2009) [148]	MC for DPMM	Both
Shi *et al.* (2011) [140]	Bayesian Yin-Yang	Both
Yang *et al.* (2012) [149]	Entropy Min.	Decremental
Iwata *et al.* (2012) [150]	MC for DPMM	Both
Fan & Bouguila (2013) [151]	Variational DP	Both
Fan & Bouguila (2014) [152]	Variational Bayes	Incremental
Kersten (2014) [153]	MML	Decremental

The primary drawback of this approach is the curse of dimensionality. For a d-dimensional problem, fitting a single full-covariance Gaussian requires $O(d^2)$ parameters, which typically forces the algorithm to restrict its models to diagonal covariances in practice. We demonstrate empirically that this approach (unsupervised learning of finite mixture models - ULFMM) does not perform well in practice, regardless of its abundant use in the literature.

Using the parsimonious factor analysis (FA) representation introduced in Section 1, it is possible to explore many models that are between a full-covariance Gaussian and a diagonal Gaussian. The resulting mixture of factor analyzers (MoFA) can be considered as a noise-robust version of the mixtures of probabilistic principal component analyzers (PPCA) approach [154]. Figure 3.3 summarizes the relations between the mixture representations in this area.

If we assume that the latent variables of each component $\mathcal{G}_k$ in a MoFA model is distributed unit normal ($\mathcal{N}(0, I)$) in the latent space, the corresponding data in the

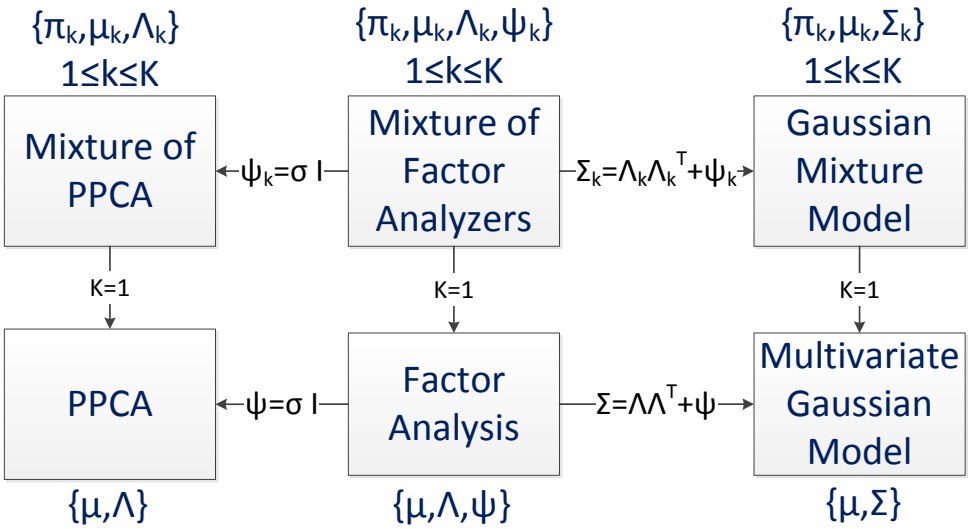

Figure 3.3. Relationship of MoFA with some well known statistical models. Model parameters are given in curly brackets. π: $(1 \times K)$ component priors, $\boldsymbol{\mu}$: $(1 \times d)$ component mean, $\mathbf{\Lambda}$: $(p \times d)$ factor loading matrix, Ψ: $(1 \times d)$ diagonal noise variances (uniqueness), $\mathbf{\Sigma}$: $(d \times d)$ component covariance. K denotes the number of components. d and p are the feature and subspace dimensionality, respectively.

feature space is also distributed as a Gaussian:

$$p\left(\mathcal{X}|\boldsymbol{z},\mathcal{G}_k\right)=\mathcal{N}\left(\boldsymbol{\mu}_k+\boldsymbol{\Lambda}_k\boldsymbol{z},\boldsymbol{\Psi}_k\right), \tag{3.12}$$

where $\boldsymbol{z}$ denotes the latent factor value. The mixture distribution of K factor analyzers is then given as [129]:

$$p(\mathcal{X})=\sum_{k=1}^{K}\int p(\mathcal{X}|\boldsymbol{z},\mathcal{G}_k)p(\boldsymbol{z}|\mathcal{G}_k)p(\mathcal{G}_k)dz. \tag{3.13}$$

The EM algorithm is used to find maximum likelihood solutions to latent variable models [128], and it can be used for training a MoFA [129]. Since EM does not address the model selection problem, it requires the number of components and factors per component to be fixed beforehand.

Ghahramani and Beal [137] have proposed a variational Bayes scheme (VBMoFA) for model selection in MoFA, which allows the local dimensionality of components and their total number to be automatically determined. In this study, we use VBMoFA as one of our benchmarking methods.

To alleviate the computational complexity of the variational approach, a greedy model selection algorithm was proposed by Salah and Alpaydın [130]. This incremental approach (IMoFA) starts by fitting a single component - single factor model to the data and adds factors and components in each iteration using fast heuristic measures until a convergence criterion is met. The algorithm allows components to have as many factors as necessary, and uses a validation set to stop model adaptation, as well as to avoid over-fitting. This is the third algorithm we use to compare with the proposed approach, which we describe in detail next.

3.2.2. Adaptive Mixtures of Factor Analyzers

We briefly summarize the proposed adaptive mixtures of factor analyzers (AMoFA) algorithm first, and then describe its details. Given a dataset $\mathcal{X}$ with N data points in d dimensions, the AMoFA algorithm is initialized by fitting a 1-component, 1-factor mixture model. Here, the factor is initialized from the leading eigenvector of the covariance matrix i. e. the principal component of the data. At each subsequent step, the algorithm considers adding more components and factors to the mixture, running EM iterations to find a parametrization. During the M-step of EM, an MML criterion is used to determine whether any weak components should be annihilated. Apart from this early component annihilation, the algorithm incorporates a second decremental scheme. When the incremental part of the algorithm no longer improves the MML criterion, a downsizing component annihilation process is initiated and all components are eliminated one by one. Similar to ULFMM, each intermediate model in both stages is stored, and the algorithm outputs the one giving the minimum message length. Figure 3.4 summarizes the proposed algorithm.

3.2.2.1. The Generalized Message Length Criterion. To allow local factor analyzers to have independent latent dimensionality, the MML criterion given in Figueiredo and Jain [143] should be generalized accordingly to reflect the individual code length of components:

$$\mathcal{L}(\boldsymbol{\theta}, \mathcal{X}) = \sum_{k:\pi_k>0} \frac{C_k}{2} \log(\frac{N\pi_k}{12}) + \frac{K_{nz}}{2} \log \frac{N}{12} + \sum_{k:\pi_k>0} \frac{(C_k+1)}{2} - \log p(\mathcal{X}|\boldsymbol{\theta}), \tag{3.14}$$

where C_k denotes the number of parameters for component k, $\mathcal{X}$ represents the dataset with N data items, $\boldsymbol{\theta}$ the model parameters, and K_{nz} represents the number of non-zero weight components. The first three terms in Equation 3.17 comprise the code length for real valued model parameters, the fourth term is the model log-likelihood, and the last two terms stand for the code length for integer hyper parameters, namely K_{nz} and

algorithm **AMoFA**(training set $\mathcal{X}$)

/*Initialization*/

$[\mathbf{\Lambda}, \boldsymbol{\mu}, \Psi] \leftarrow$ train a 1-component, 1-factor model

repeat

/*Perform a single split*/

x $\leftarrow$ Select a component for splitting via Equation 3.18

$[\mathbf{\Lambda}_1, \boldsymbol{\mu}_1, \Psi_1, \pi_1] \leftarrow$ MML_EM(split x).

actionML(1) $\leftarrow$ ML($\mathbf{\Lambda}_1, \boldsymbol{\mu}_1, \Psi_1, \pi_1$) via Equation 3.17

/*Perform a single factor addition*/

y $\leftarrow$ Select a component to add a factor

$[\mathbf{\Lambda}_2, \boldsymbol{\mu}_2, \Psi_2, \pi_2] \leftarrow$ MML_EM(add factor to y).

actionML(2) $\leftarrow$ ML($\mathbf{\Lambda}_2, \boldsymbol{\mu}_2, \Psi_2, \pi_2$) via Equation 3.17

/*Select the best action*/

z $\leftarrow$ arg min(actionML(1),actionML(2))

/*Update the parameters*/

$[\mathbf{\Lambda}, \boldsymbol{\mu}, \Psi, \pi] \leftarrow [\mathbf{\Lambda}_z, \boldsymbol{\mu}_z, \Psi_z, \pi_z]$

until MML decrease $< \epsilon$

/*Annihilation starts with $k = K$ components*/

while $k > 1$

/*Select the weakest component for annihilation*/

$[\mathbf{\Lambda}_k, \boldsymbol{\mu}_k, \Psi_k, \pi_k] \leftarrow$ EM(annihilate component).

k = k - 1

end

/*Select l that minimizes MML criterion in Equation 3.17*/

return $[\mathbf{\Lambda}_l, \boldsymbol{\mu}_l, \Psi_l, \pi_l]$

end

Figure 3.4. Outline of the AMoFA algorithm.

component-wise latent dimensionalities $\{p_k\}$. We use here Rissanen's *universal prior for integers* [134]:

$$w^*(k) = c^{-1}2^{-log^*k}, \tag{3.15}$$

which gives the (ideal) code length

$$L^*(k) = log\ 1/w^*(k) = log^*(k) + log\ c, \tag{3.16}$$

where $log^*(k) = logk + loglogk + ...$ is n-fold logarithmic sum with positive terms, c is the normalizing sum $\sum_{k>0} 2^{-log^*k}$ that is tightly approximated as $c = 2.865064$ [134]. $log^*(k)$ term in Equation 3.16 can be computed via a recursive algorithm. We finally obtain $L^*(K_{nz})$, the cost to encode the number of components, and similarly $\sum_{k:\pi_k>0} L^*(p_k)$, the cost to encode the local dimensionalities $\{p_k\}$ and add them to Equation (3.14) to obtain a message length criterion:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \mathcal{X}) = &\sum_{k:\pi_k>0} \frac{C_k}{2} \log(\frac{N\pi_k}{12}) + \frac{K_{nz}}{2} \log \frac{N}{12} \\ &+ \sum_{k:\pi_k>0} \frac{(C_k+1)}{2} - \log p(\mathcal{X}|\boldsymbol{\theta}) \\ &+ L^*(K_{nz}) + \sum_{k:\pi_k>0} L^*(p_k)\end{aligned} \tag{3.17}$$

3.2.2.2. Component Splitting and Factor Addition. Adding a new component by splitting an existing one involves two decisions: which component to split, and how to split it. AMoFA splits the component that looks least likely to a Gaussian, by looking at a multivariate kurtosis metric [155]. For a multinormal distribution, the multivariate kurtosis takes the value $\beta_{2,d} = d(d+2)$, and if the underlying population is multivariate normal with mean $\boldsymbol{\mu}$, the sample counterpart of $\beta_{2,d}$, namely $b_{2,d}$, has an asymptotic distribution as the number of samples N goes to infinity. Salah and Alpaydın [130]

adapted this metric to the mixture model by using a "soft count" $h_j^t \equiv E[\mathcal{G}_j|\boldsymbol{x}^t]$:

$$\gamma_j = \{b_{2,d}^j - d(d+2)\}\left[\frac{8d(d+2)}{\sum_{t=1}^N h_j^t}\right]^{-\frac{1}{2}} \tag{3.18}$$

$$b_{2,d}^j = \frac{1}{\sum_{l=1}^N h_j^l}\sum_{t=1}^N h_j^t\left[(\boldsymbol{x}^t - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}^t - \boldsymbol{\mu}_j)\right]^2 \tag{3.19}$$

The component with greatest γ_j is selected for splitting. AMoFA runs a local, 2-component MoFA on the data points that fall under the component. To initialize the means of new components prior to MoFA fitting, AMoFA uses the weighted sum of all eigenvectors of the local covariance matrix: $\boldsymbol{w} = \sum_i^d \boldsymbol{v}_i\lambda_i$, and set $\boldsymbol{\mu}_{new} = \boldsymbol{\mu} \pm \boldsymbol{w}$, where $\boldsymbol{\mu}$ is the mean vector of the component to split.

The component having the largest difference between modeled and sample covariance is selected for factor addition. As in IMoFA, AMoFA uses the residual factor addition scheme. Given a component $\mathcal{G}_j$ and a set of data points $\boldsymbol{x}^t$ under it, the re-estimated points after projection to the latent subspace can be written as: $\tilde{\boldsymbol{x}}_j^t = \boldsymbol{\Lambda}_j E[\boldsymbol{z}^t|\boldsymbol{x}^t, \mathcal{G}_j]$. The re-estimation error decreases with the number of factors used in $\boldsymbol{\Lambda}_j$. The newly added column in the factor loading matrix, $\boldsymbol{\Lambda}_{j,p+1}$, is selected to be the principal direction (the eigenvector with the largest eigenvalue) of the residual vectors $\tilde{\boldsymbol{x}}_j^t - \boldsymbol{x}_j^t$. This new factor is used in bootstrapping the EM procedure.

3.2.2.3. Component Annihilation. In a Bayesian view, the message length criterion (Equation 3.17) adapted from Figueiredo and Jain [143] corresponds to assuming a flat prior on component parameters $\boldsymbol{\theta}_k$, and a Dirichlet prior on mixture proportions π_k:

$$p(\pi_1, \cdots, \pi_K) \propto exp\{\sum_{k=1}^{K_{nz}} -\frac{C_k}{2} log\pi_k\} = \prod_{k=1}^{K_{nz}} \pi_k^{-C_k/2}. \tag{3.20}$$

Thus, in order to minimize the adapted cost in Equation 3.17, the M-step of EM is changed for π_k :

$$\hat{\pi}_k^{new} = \frac{max\{0, (\sum_{i=1}^{N} h_{ik}) - \frac{C_k}{2}\}}{\sum_{j=1}^{K_{nz}} max\{0, (\sum_{i=1}^{N} h_{ij}) - \frac{C_k}{2}\}}, \tag{3.21}$$

which means that all components having a soft count (N_k) smaller than half the number of local parameters C_k will be annihilated. This threshold enables the algorithm to get rid of components that do not justify their existence. In the special case of AMoFA, the number of parameters per component are defined as:

$$C_k = d * (p_k + 2) + L^*(p_k), \tag{3.22}$$

where d is the original dataset dimensionality, p_k is the local latent dimensionality of component k, and $L^*(p_k)$ is the code length for p_k. The additive constant 2 inside the bracket accounts for the parameter cost of mean $\boldsymbol{\mu}_k$ and local diagonal uniquenesses matrix $\boldsymbol{\Psi}_k$. Finally, the localized annihilation condition to check at the M step of EM is simply $N_k < T_k = C_k/2$.

In AMoFA, we use an outer loop to drive the model class adaptation and an inner EM loop to fit a mixture of factor analyzer model with initialized parameters. The inner EM algorithm is an improved and more generalized version of ULFMM [143], where after parallel EM updates we select the weakest component and check $N_k < T_k$ for annihilation, as opposed to sequential component update approach (using Component-wise EM -CEM^2 [156]). Any time during EM, automatic component annihilation may take place. When the incremental progress is saturated, the downsizing component annihilation is initiated. The MML based EM algorithm and relevant details are given in the next section.

3.2.2.4. EM Algorithm for Mixture of Factor Analyzers with MML Criterion. In this subsection, we introduce the MoFA EM algorithm optimizing the generalized MML criterion given in the former subsection. This criterion is used for automatic annihi-

lation of components at the M step. We first provide the formulation of regular EM algorithm for MoFA model [129]:

$$
\begin{aligned}
E[\boldsymbol{z}|\mathcal{G}_k, \boldsymbol{x}^t] &= h_{ik}\Omega_k \left(\boldsymbol{x}^t - \boldsymbol{\mu}_k\right) && (3.23)\\
E\left[\boldsymbol{z}\boldsymbol{z}'|\mathcal{G}_k, \boldsymbol{x}^t\right] &= h_{ik}(I - \Omega_k\boldsymbol{\Lambda}_k + \Omega_k(\boldsymbol{x}^t - \boldsymbol{\mu}_k)(\boldsymbol{x}^t - \boldsymbol{\mu}_k)'\Omega') && (3.24)\\
\tilde{\boldsymbol{\Lambda}}_k^{\text{new}} &= \left(\sum_i h_{ik}\boldsymbol{x}^t E\left[\tilde{\boldsymbol{z}}|\boldsymbol{x}^t, \mathcal{G}_k\right]'\right)\left(\sum_j h_{jk}E[\boldsymbol{z}\boldsymbol{z}'|x_j, \mathcal{G}_k]\right)^{-1} && (3.25)\\
\boldsymbol{\Psi}_k^{new} &= \frac{1}{N\pi_k} diag\{\sum_i h_{ik}(\boldsymbol{x}^t - \tilde{\boldsymbol{\Lambda}}_k^{new}E[\tilde{\boldsymbol{z}}|\boldsymbol{x}^t, \mathcal{G}_k])\boldsymbol{x}^{t\prime}\} && (3.26)\\
\pi_k^{\text{new}} &= \frac{1}{N}\sum_{i=1}^{N} h_{ik} && (3.27)
\end{aligned}
$$

where to keep the notation uncluttered, $\tilde{\boldsymbol{z}}$ is defined as $\begin{bmatrix}\boldsymbol{z} & 1\end{bmatrix}'$. Similarly, $\tilde{\boldsymbol{\Lambda}}_k = \begin{bmatrix}\boldsymbol{\Lambda}_k & \boldsymbol{\mu}_k\end{bmatrix}$, $\Omega_k \equiv \boldsymbol{\Lambda}_k(\boldsymbol{\Psi}_k + \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k{}')^{-1}$, and

$$
h_{ik} = E\left[\mathcal{G}_k|\boldsymbol{x}^t\right] \propto p(\boldsymbol{x}^t, \mathcal{G}_k) = \pi_k\mathcal{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k{}' + \boldsymbol{\Psi}_k\right). \tag{3.28}
$$

The above EM formulation is optimizing the MoFA log likelihood, which is the logarithm of the linear combination of component likelihoods:

$$
\begin{aligned}
p\left(\mathcal{X}|\boldsymbol{z}, \mathcal{G}\right) &= log\prod_{i=1}^{N}\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\boldsymbol{x}^t; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k{}' + \boldsymbol{\Psi}_k\right)\\
&= \sum_{i=1}^{N} log\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k{}' + \boldsymbol{\Psi}_k\right).
\end{aligned} \tag{3.29}
$$

The EM Algorithm for MoFA using MML criterion is given in Figure 3.5. Note that the automatic component annihilation uses a local threshold derived from the MML criterion. Note also that unlike Figueiredo and Jain, AMoFA does not use Component-wise EM algorithm (CEM^2) [156] for component update and annihilation. The reason Figueiredo and Jain use CEM^2 is that when starting with a large K_{max}, immature components may not have the necessary support to survive. Nevertheless, using a sequential annihilation scheme bears the risk of losing a well-fitting

component with slightly less support than the threshold. To overcome this problem, AMoFA uses regular EM updates for all components, then proceeds with annihilating the weakest component below threshold. Besides, since its adaptive approach is incremental, AMoFA typically does not generate a multitude of weak components at a time. Usually, the components die in time when they lose support.

Since the maximum likelihood estimation via the EM algorithm generally leads parameters to the boundary of the parameter space, it is common to use covariance regularization [149,157], no matter whether a penalty term is used to tune model complexity or not. In AMoFA, the regularization term is added directly to the uniquenesses matrix:

$$\mathbf{\Psi}_k^{reg} = (1-\gamma)\mathbf{\Psi}_k + \gamma \sum_{l=1} \mathbf{\Psi}_l \tag{3.30}$$

where γ is a small constant (e. g. 10^{-4}) for regularization. Within the algorithm, AMoFA also compensates for the changes of modeled feature variances when factors are added. This is achieved by removing the diagonal responsibility of the changed factor loading vectors from the uniquenesses matrix. When a new factor is added to component k we have:

$$\Psi_{kj}^{reg} \quad = \quad \Psi_{kj} - Ł_{kj}^2, \tag{3.31}$$

where $Ł_{kj}$ is the j^{th} entry of new factor loading vector.

3.2.3. Illustration of Automatic Mixture Model Selection on Synthetic Data

In this section, some examples on benchmarking clustering problems are given to illustrate the functioning of AMoFA algorithm as well as to compare its performance with closely related model selection methods.

3.2.3.1. Evaluation Protocol for Clustering Performance. We use the Normalized Information Distance (NID) metric for evaluating clustering accuracy, as it possesses several important properties; in addition to being a metric, it admits an analytical adjustment for chance, and allows normalization to [0-1] range [158]. NID is formulated as:

$$1 - \frac{I(\boldsymbol{u}, \boldsymbol{v})}{\max\{H(\boldsymbol{u}), H(\boldsymbol{v})\}}, \tag{3.32}$$

where entropy $H(\boldsymbol{u})$ and the mutual information $I(\boldsymbol{u}, \boldsymbol{v})$ for clustering are defined as follows:

$$H(\boldsymbol{u}) = -\sum_{i=1}^{R} \frac{a_i}{N} \log \frac{a_i}{N}, \tag{3.33}$$

$$I(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{a_i b_j / N^2}, \tag{3.34}$$

Here, a_i is the number of samples in cluster i, n_{ij} is the number of samples falling into cluster i in clustering $\boldsymbol{u}$ and cluster j in clustering $\boldsymbol{v}$. MI is a nonlinear measure of dependence between two random variables. It quantifies how much information in bits the two variables share. We compute NID between the ground truth and the clusterings obtained by the automatic model selection techniques in order to give a more precise measure of clustering than just the number of clusters. When there is no overlap, NID is expected to be close to 0; higher overlap of clusters might result in higher average NID, although a relative performance comparison can still be achieved.

3.2.3.2. Experiments on Benchmark Datasets for Clustering. We tested three methods, namely IMoFA [130], AMoFA and ULFMM [143] on benchmark synthetic/real datasets for clustering. For maximum comparability with previous work, we used some sythetic dataset examples from Figueiredo and Jain [143] as well as from a recent study on automatic mixture model selection [149].

AMoFA, as opposed to IMoFA and ULFMM, does not rely on random initial-

ization. In IMoFA, the first factor is randomly initialized, and in ULFMM the initial cluster centers are assigned to randomly selected instances. AMoFA, on the other hand, initializes the first factor from the principal component of the dataset. Similar to residual factor addition, this scheme can be shown to converge faster than random initializations. Given a dataset, a single simulation is sufficient to assess performance.

Because of this deterministic property of AMoFA, we report the results with multiple datasets sampled from the underlying distribution, instead of sampling once and simulating multiple times. Unless stated otherwise, in the following experiments with synthetic datasets, 100 samples are drawn and the average results are reported. For ULFMM, we give initial number of clusters $K^{max} = 20$ in all our simulations for clustering and use free full covariances. Moreover, the EM convergence threshold ϵ is set to 10^{-5} in all three methods.

Example 1: 3 Separable Gaussians. As a toy example, we generated a mixture of three Gaussians having the same mixture proportions $\pi_1 = \pi_2 = \pi_3 = 1/3$ and the same covariance matrix $diag\{2, 0.2\}$ with separate means $\boldsymbol{\mu}_1 = [0, -2]', \boldsymbol{\mu}_2 = [0, 0]', \boldsymbol{\mu}_3 = [0, 2]'$. Different from previous work, where this dataset had been used [143, 149] we test the synthetic data with 900 data points. We generate 100 samples from the underlying distribution.

Figure 3.6 shows the evolution of adaptive steps of AMoFA with found clusters shown in 2-std contour plots, and the description length (DL) is given above each plot. To keep the figure uncluttered, only the mixture models obtained at the end of adaptive steps are given. The initial step fits a single component-single factor model. The first two iterations add components to the mixture, and the next one adds a factor. The incremental phase stops when no improvement in the message length is observed. Then, the algorithm starts to annihilate the components, until a single component is left. The DL in the decremental phase is higher, since components have two factors. Finally, the algorithm selects the 3-component solution having the minimum DL.

Require: $\mathcal{X}$ data, and initialized MoFA parameter set $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \pi\}$

REPEAT

E Step: compute expectations $h_{ik}, \mathrm{E}[\boldsymbol{z}|\mathcal{G}_k, \boldsymbol{x}^t], E\left[\boldsymbol{z}\boldsymbol{z}'|\mathcal{G}_k, \boldsymbol{x}^t\right]$ using Equation 3.28, 3.23 and 3.24, respectively

M step: compute model parameters using Equations 3.25-3.27

Compute $T_k = C_k/2$ using Equation 3.22

while any component needs annihilation

Annihilate **the weakest** component k having $N_k < T_k$

Update $\pi_k = \pi_k / \sum_{l=1}^{K_{nz}^{new}} \pi_l, 1 \leq k \leq K_{nz}^{new}$

end

Compute message length $\mathcal{L}(\boldsymbol{\theta}, \mathcal{X})$ using Equation 3.17

UNTIL $\mathcal{L}(\boldsymbol{\theta}, \mathcal{X})$ converges with ϵ tolerance

Figure 3.5. EM Algorithm for MoFA with MML Criterion.

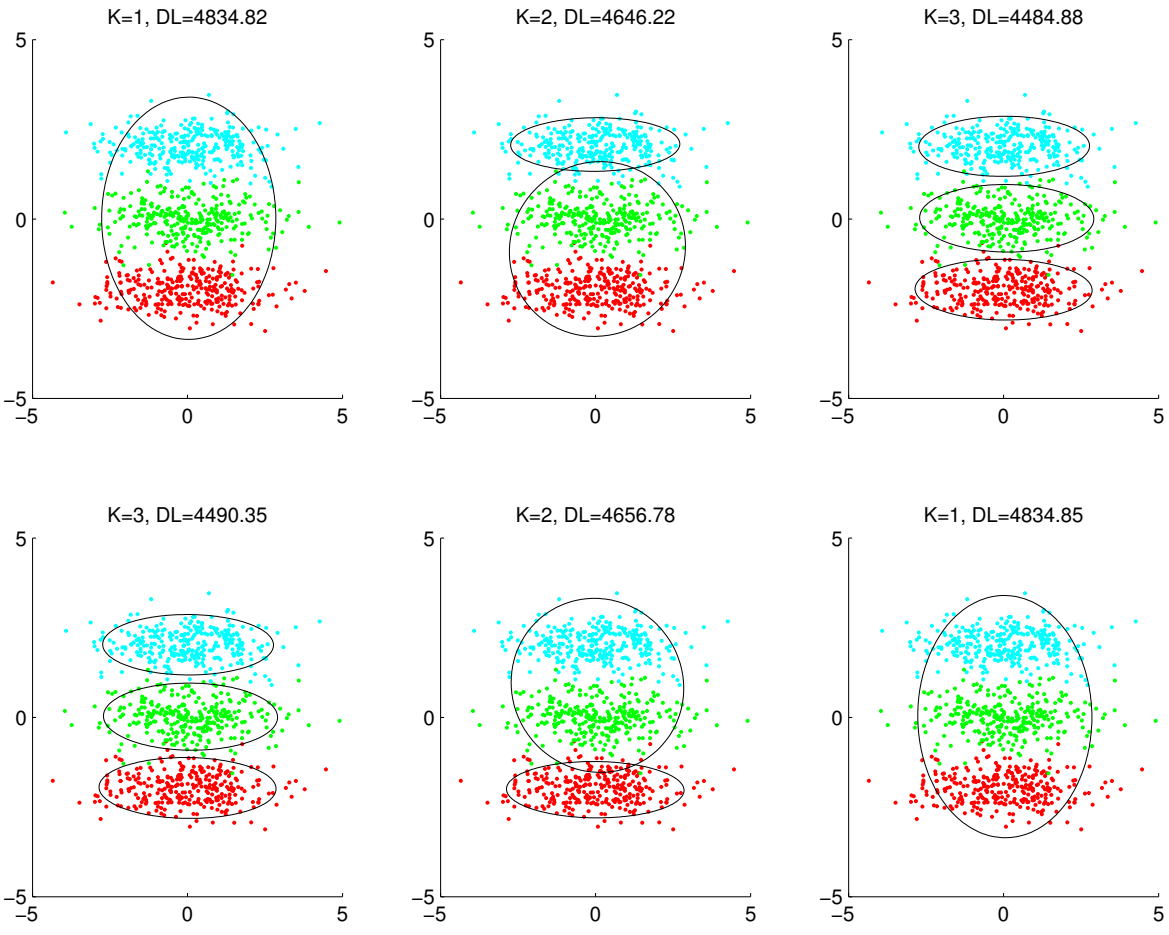

Figure 3.6. The evolution of AMoFA on a toy synthetic data.

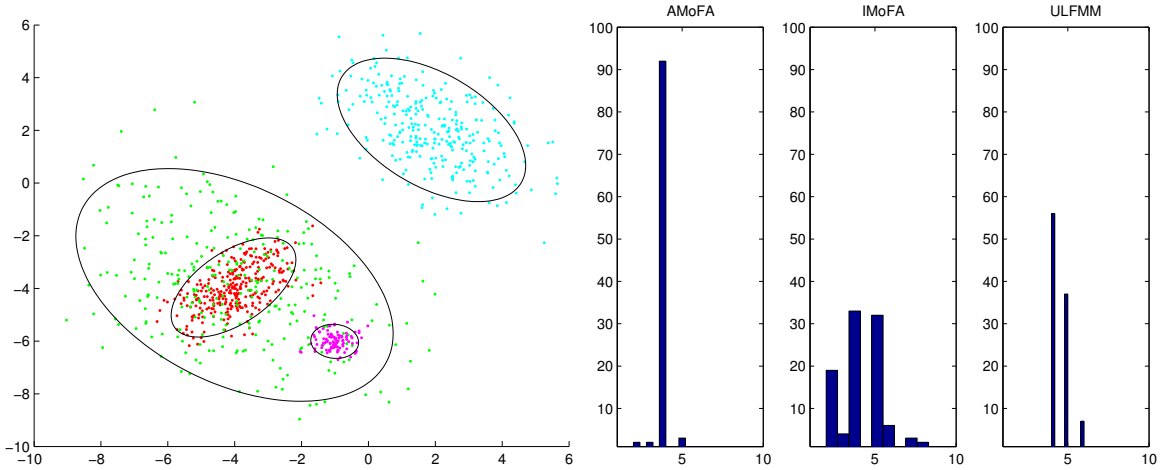

Figure 3.7. Overlapping Gaussians data. Left: A sample AMoFA result. The real labels are shown with colors and resulting AMoFA mixture model is shown with 2-std contour plot. Right: Histograms of number of clusters found by AMoFA, IMoFA and ULFMM respectively.

Example 2: Overlapping Gaussians. As a more challenging clustering task, we use an example very similar to the one used in [143, 149]. Here, three of the four Gaussians overlap with the following generative model:

$\pi_1 = \pi_2 = \pi_3 = 0.3,\ \pi_4 = 0.1,$

$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = [-4\ \ -4]',\ \boldsymbol{\mu}_3 = [2\ 2]',\ \boldsymbol{\mu}_4 = [-1\ \ -6]',$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} .8 & .5 \\ .5 & .8 \end{bmatrix},\ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix},\ \boldsymbol{\Sigma}_3 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix},\ \boldsymbol{\Sigma}_4 = \begin{bmatrix} 0.125 & 0 \\ 0 & 0.125 \end{bmatrix}.$$

We use $N = 1000$ data points. As in the previous example, we generate 100 random datasets. In Figure 3.7 left plot, the data are illustrated with a sample result of AMoFA. Out of 100 simulations, the accuracy of finding K*=4 is 92, 56, and 33 for AMoFA, ULFMM, and IMoFA, respectively. The histogram in Figure 3.7 right plot shows the distribution of number of automatically found clusters for three methods. Average NID over 100 datasets is found to be 0.2549, 0.2951, and 0.3377 for AMoFA, ULFMM and IMoFA, respectively. A paired t-test (two tailed) on NID scores indicates that AMoFA performs significantly better than ULFMM with $p < 10^{-5}$.

Example 3: Geva's Face This example is originally used by Geva [159] to imitate a face using a mixture of five Gaussians. We use the following generative model [149]

to generate 100 datasets, each with 1000 number of instances:
$\pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = 0.2$,
$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = [0\ 0]',\ \boldsymbol{\mu}_3 = [-1.5\ 1.5], \boldsymbol{\mu}_4 = [1.5\ 1.5],\ \boldsymbol{\mu}_5 = [0\ 1.5]$,
$\boldsymbol{\Sigma}_1 = diag\{0.01, 1.25\}, \boldsymbol{\Sigma}_2 = diag\{8, 8\}, \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_4 = diag\{0.02, 0.015\}, \boldsymbol{\Sigma}_5 = diag\{1, 0.2\}$

A sample progress of AMoFA on Geva's Face data is given in Figure 3.8. Using an entropy minimization based model selection method, Yang *et al.* report 90% correct clustering accuracy out of 100 generated datasets [149]. We obtain 98% with ULFMM, 98% with AMoFA, and 68% with IMoFA, respectively.

3.2.4. Application to Classification: Modeling Class Conditional Densities

In addition to ULFMM [143] and the IMoFA [130], we compare AMoFA also with VBMoFA [137] on classification. As baseline, we use Mixture of Gaussians, where the data of each class are modeled by a single Gaussian with full (MoG-F) or diagonal (MoG-D) covariances. We compare the performances of the methods via class-conditional modeling. on nine benchmark datasets: The ORL face database with binary gender classification task [160], 16-class phoneme database from LVQ package of Helsinki University of Technology [161], the VISTEX texture database [130], a 6-class Japanese Phoneme database [162], the MNIST dataset [163], and four datasets (Letter, Pendigits, Opdigits, and Waveform) from UCI ML Repository [164]. Table 3.2 gives some basic statistics about the databases. Except MNIST that has an explicit train and testing protocol, all experiments were carried out with 10-fold cross-validation. Simulations are replicated 10 times in MNIST, where we crop the 4 pixel padding around the images then scale them to 10x10 pixels and obtain feature vectors $\boldsymbol{x} \in \mathbb{R}^{100}$.

In the experiments, we trained separate mixture models for the samples of each class, and used maximum likelihood classification. We did not use informative class priors, as it would positively bias the results, and hide the impact of likelihood modeling. In Table 3.3, we provide accuracy computed over 10 folds, where all four approaches used the same protocol. ULFMM column reports performance of ULFMM models with free diagonal covariances, as full covariance models invariably give poorer results.

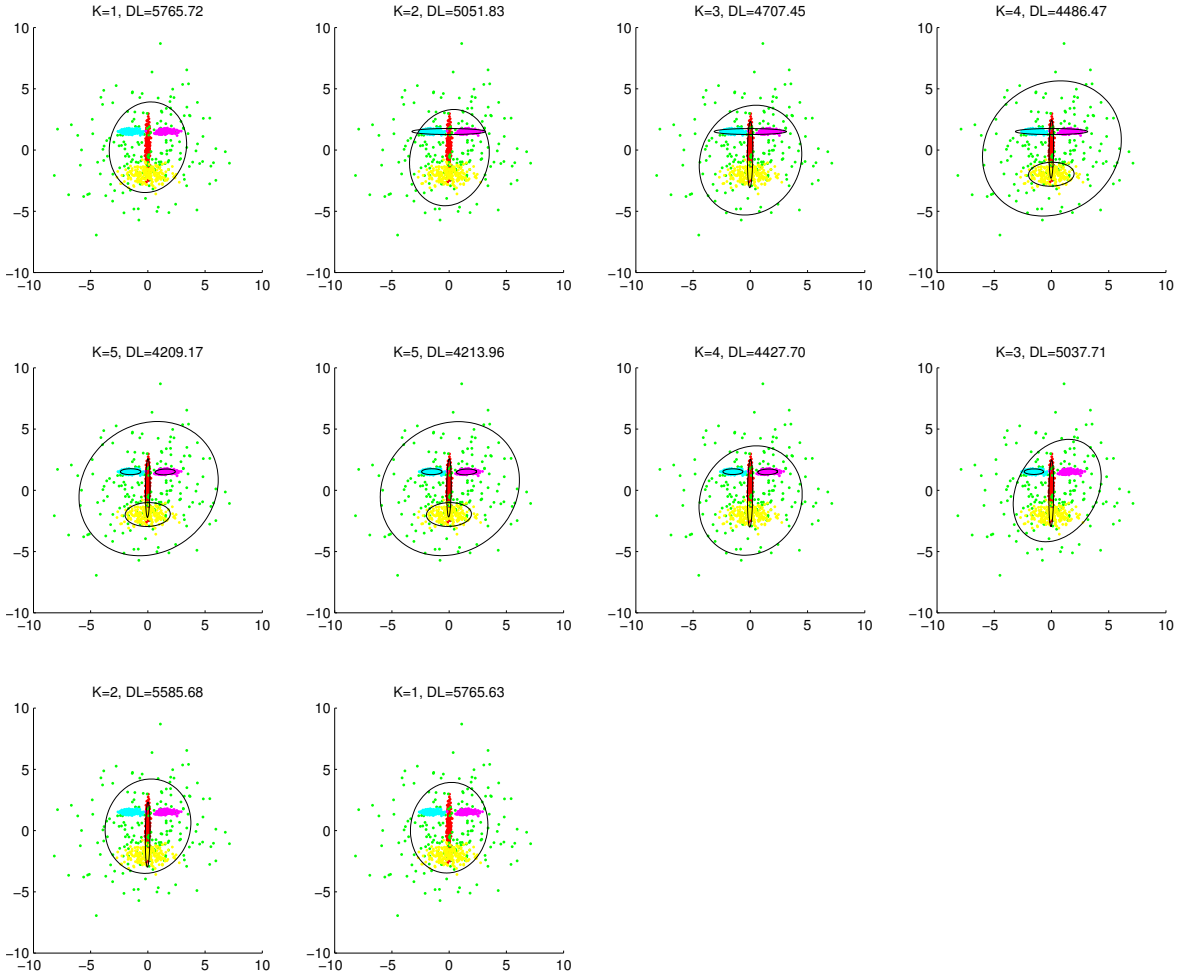

Figure 3.8. Evolution of AMoFA for face imitation data. We see that the facial features emerge incrementally via component splitting and correctly locating clusters. AMoFA does not find further splitting more feasible and starts factor addition, which results in increases in mixture DL. Finally, the progress gets saturated and downsizing component annihilation takes place.

Table 3.2. Datasets used for class conditional mixture modeling.

Dataset	Dimensions	Classes	# of Samples
ORL	256	2	400
LVQ	20	16	3858
OPT	60	10	4677
PEN	16	10	8992
VIS	169	10	3610
WAV	21	3	500
JPN	112	6	1200
LET	16	26	20000
MNT	100	10	70000

The best results for a dataset are shown in **bold**. We compared the algorithms with a non-parametric sign test. For each dataset, we conducted a one tail paired-sample t-test with a significance level of 0.05 (0.01 upon of rejection of null hypothesis). Results indicate that ULFMM ranks the last in all cases: even against MoG-F baseline it is always inferior. This is because of the fact that after randomized initialization of clusters, ULFMM algorithm annihilates all illegitimate components skipping intermediate (possibly better than initial) models. On seven datasets AMoFA attains/shares the first rank, and on the remaining two it ranks the second. Note that although on the overall AMoFA and VBMoFA have similar number of wins against each other, on high dimensional datasets, namely on MNIST, VISTEX, Japanese Phoneme and ORL, AMoFA significantly outperforms VBMoFA. The results of pairwise tests at 0.05 significance level are shown in Table 3.4. We see that the adaptive MoFA algorithms dramatically outperform GMM based ULFMM algorithm. MoFA is capable of exploring a wider range of models between diagonal and full covariance with reduced parameterization. Among the three MoFA based algorithms, no significant difference ($\alpha = 0.05$) was found on Pendigits dataset. AMoFA outperforms the best results reported so far with the VISTEX dataset. The best test set accuracy reported in [130] is 73.8 ± 1.1 using GMMs. We attain 77.2 ± 4.6 with AMoFA.

Table 3.3. Classification accuracies for class-conditional models. Significantly better results compared to the first runner up are shown in **bold**, where * signifies 0.05 significance level, while ** corresponds to 0.01 significance level. If there are multiple best performers without pair-wise significant difference, they are shown in bold altogether.

	IMoFA-L [130]	VBMoFA [137]	AMoFA
ORL	**97.8 ± 1.5**	93.0 ± 2.8	**97.5 ± 1.2**
LVQ	**91.2 ± 1.9**	**91.3 ± 1.9**	89.3 ± 1.6
OPT	91.1 ± 2.7	**95.2 ± 1.8**	93.8 ± 2.4
PEN	**97.9 ± 0.7**	**97.8 ± 0.6**	**98.1 ± 0.6**
VIS	69.3 ± 4.6	67.1 ± 5.9	**77.2 ± 4.6****
WAV	80.8 ± 4.5	**85.1 ± 4.2**	**85.6 ± 4.6**
JPN	93.4 ± 2.4	93.2 ± 3.2	**96.5 ± 2.2***
LET	86.6 ± 1.5	**95.2 ± 0.7**	**95.1 ± 0.7**
MNT	91.5 ± 0.2	84.5 ± 0.1	**93.9 ± 0**
	ULFMM [143]	MoG-D	MoG-F
ORL	80.0 ± 6.5	89 ± 2.4	90 ± 0
LVQ	75.4 ± 4.5	88.1 ± 2.6	**92.1 ± 1.8**
OPT	49.5 ± 10.2	84.2 ± 3.1	**94.9 ± 1.7**
PEN	89.9 ± 2.0	84.5 ± 2.0	97.4 ± 0.6
VIS	20.6 ± 3.7	68.6 ± 3.9	44.7 ± 12.8
WAV	72.1 ± 7.5	80.9 ± 18.2	**84.8 ± 4.6**
JPN	82.4 ± 2.1	82.2 ± 4.9	92.3 ± 2.3
LET	56.9 ± 2.8	64.2 ± 1.2	88.6 ± 0.9
MNT	64.7 ± 2.0	78.2 ± 0	**93.7 ± 0**

Table 3.4. Pairwise wins/ties/losses with 0.05 significance.

	AMoFA	VBMoFA	ULFMM	MoG-D	MoG-F
IMoFA	1/2/6	2/4/3	9/0/0	7/2/0	2/3/4
AMoFA	*	4/3/2	9/0/0	7/2/0	5/3/1
VBMoFA		*	9/0/0	7/2/0	3/5/1
ULFMM			*	0/3/6	0/0/9
MoG-D				*	1/2/6

3.2.5. Application to Acoustic Emotion Recognition in Natural Conditions

We apply the AMoFA algorithm to the challenging task of acoustic emotion recognition in-the-wild using a recently introduced emotional child Russian speech corpus (EmoChildRU) [165]. EmoChildRU is a subset of ChildRU [166], that features over 20K utterances (30 hours of recording) of 100 children (aged 3-7 years) in their natural environment (e. g. home, orphanage or kindergarten) while doing a set of activities (playing with toys, talking to their mother/teacher or a child psychologist, watching a Russian cartoon etc.). The subset we are interested here is annotated for three affective valence related states, namely comfort, discomfort and neutral, by five child speech specialists using audio-visual and linguistic cues. It is known that valence classification from acoustics is more difficult compared to arousal. Moreover, the recordings are in natural conditions, where the child's distance to microphone and the acoustic recording environment vary markedly on top of speaker and linguistic content variability. On the overall, the conditions resemble a cross corpus setting where the results reported without corpus normalization strategies are marginally above chance level [3].

We split given 585 instances into speaker disjoint training and test sets as shown in Table 3.5. As a baseline we extract suprasegmental openSMILE [167] features with the configuration file used in the latest INSTERSPEECH ComParE challenges since 2013 [4]. For AMoFA, we also use openSMILE to extract LLDs with a configuration from INTERSPEECH 2010 Challenge [168]. This LLD set includes 38 popularly used LLDs and their first order temporal derivatives giving 76 dimensional features.

For our application, we employ class-conditional AMoFA modeling of LLDs. When a test utterance is obtained, its LLDs are given to three class conditional models and the class giving the highest mixture likelihood is taken as prediction. For the baseline system, we use Linear Kernel SVM with default parameter setting (complexity parameter set to one) to train 6 373 dimensional suprasegmental features. Note that this classifier-feature set combination is used in recent ComParE challenges [4, 6, 169].

Table 3.5. Distribution of classes and gender (M/F) in train and test partitions.

	M/F	Comfort	Neutral	Discomfort	Total
Train	16/20	144	164	52	**360**
Test	7/7	90	88	47	**225**
Total	**23/27**	**234**	**252**	**99**	**585**

We report classification performance in terms of Unweighted Average Recall (UAR), which is used for the first time in INTERSPEECH 2009 Emotion Recognition Challenge [65], in order to avoid bias towards the majority class. When learning fails, UAR takes the chance level value of $1/K$, where K is the number of classes. In our case, the chance level UAR is $1/3 = 33.33\%$. The comparative results of baseline systems and AMoFA are given in Table 3.6. We see that without normalization, suprasegmental openSMILE features fail totally (see the first row in Table 3.6). The UAR performance of the baseline system reaches 45.3% and 45.7%, with z-norm and min-max norm, respectively. Without any normalization or need for parameter optimization, AMoFA gives an UAR of 46.4%. As mentioned earlier, the low overall performance stems from the challenging nature of the data and the classification task.

Table 3.6. UAR (%) performance comparison of baseline SVM systems and AMoFA system on test set of EmoChildRU.

Classifier	Feature Set	Normalization	UAR
SVM	IS13-Func	No-norm	34.1
SVM	IS13-Func	Z-norm	45.3
SVM	IS13-Func	Minmax	45.7
AMoFA	IS10-LLD	No-norm	46.4

3.2.6. Overview

In this chapter, we propose a novel and adaptive model selection approach for Mixtures of Factor Analyzers. Our algorithm first adds factors and components to the mixture, and then prunes excessive parameters, thus obtaining a parsimonious model in a very time and space efficient way. Our contributions include a generalization of the adapted MML criterion to reflect local parameter costs, as well as local component annihilation thresholds.

We carry out experiments on synthetic and real datasets, and the results indicate the superiority of the proposed method in clustering and class-conditional modeling. We contrast our approach with the Incremental MoFA approach [130], Variational Bayesian MoFA [137], as well as the popular ULFMM algorithm [143].

In high dimensions, MoFA based automatic modeling provides significantly better classification results than GMM based ULFMM modeling, as it is capable of modeling a much wider range of models with compact parametrization. It also makes use of the latent dimensionality of the local manifold, thus enables obtaining an adaptive cost for the description length. AMoFA algorithm is observed to offer the best performance on higher dimensional datasets.

An application to acoustic emotion recognition shows that AMoFA outperforms the state-of-the-art system based on SVM and suprasegmental openSMILE features. Application to other in-the-wild affective and paralinguistic tasks is left for future work.

3.3. Links and Future Research Directions

In this chapter, we dealt with FA and CCA as unrelated statistical methods. Here we will draw the link and mention how this relationship can be exploited in future works.

We know that FA and Probabilistic PCA (PPCA) are closely related, and that

they only differ in the way they model the noise variance. We also know that CCA degenerates to PCA when one of the two views is the identity matrix, and to LDA when one view is one-of-K coded target matrix. The relationship of CCA and FA lies in the probabilistic interpretation of CCA, namely PCCA [170]. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote observed random variables of two views/representations of the latent variable $\boldsymbol{z}$. The probabilistic graphical model of this scheme is the same as that of factor analysis with $\boldsymbol{v} = (\boldsymbol{x}, \boldsymbol{y})$ as observed variable. Therefore, PCCA is nothing but FA applied on the combination of the two views.

This relationship can be exploited especially when we are interested in mixture extension of PCCA (MPCCA). Therefore, MPCCA models can be effectively learned via MoFA. Furthermore, benefiting from the relationship of CCA and LDA, MoFA can also be employed as mixtures of probabilistic LDA (MPLDA). Note that, CCA itself can be used in supervised, unsupervised as well as semi-supervised settings [100, 118, 119]. As long as the contributions of this thesis are concerned, (i) we can initialize multi-view AMoFA with traditional CCA instead of PCA, and (ii) we can employ AMoFA for multimodal extension of the CCA based feature selection methods introduced in Section 3.1. Moreover, AMoFA automatically resolves the model selection issue for other MPCCA applications, such as the multi-view super vector introduced for video modeling [171].

4. APPLICATIONS IN COMPUTATIONAL PARALINGUISTICS

In this chapter, we illustrate the performance of the proposed methods on a set of paralinguistic applications. As mentioned earlier, the data are from recent challenges, and all experiments adhere to the standard challenge protocol. We provide experimental results for the proposed methods on conflict recognition (INTERSPEECH ComParE 2013), physical load recognition (INTERSPEECH ComParE 2014), and eating condition recognition (INTERSPEECH ComParE 2015).

For conflict and physical load binary classification tasks, we employ two of the proposed feature selection (FS) methods, namely SLCCA-RAND [11] and SLCCA-LLD [10], respectively. These are extended versions of SLCCA-Filter [9], to overcome the issue of covariance singularity therefore the curse of dimensionality. Although both approaches use SLCCA-Filter as base ranker and employ a divide-and-conquer strategy they differ in the final evaluation of the base rankings. SLCCA-RAND aggregates the feature weights over iterations of randomly partitioned subsets and obtains a single saliency vector for a final ranking. In contrast, SLCCA-LLD applies SLCCA-Filter to domain-knowledge inspired feature groups and concatenates top ranking features from each group. This approach not only increases the accuracy of FS, but also reduces the space/time complexity, dramatically. SLCCA-RAND is proposed after the respective challenge, giving the state-of-the-art results on the test set of the Conflict Corpus [11], with a much more efficient FS approach compared to the one used by the challenge winner [79]. SLCCA-LLD is proposed in the paper participated in the INTERSPEECH 2014 ComParE Challenge, and won the respective Sub-Challenge.

These two FS approaches are also employed in the eating condition challenge. However, their contribution on the baseline feature set was incremental compared to the cascaded normalization scheme, which includes speaker and non-linear normalization. Therefore, we decided on a paradigm change and employed the Fisher Vector (FV)

encoding of Low Level Descriptors (LLD) to alleviate the speaker variability. The FV is a powerful approach that is recently popularly used in computer vision. We employ the FV encoding for a paralinguistic task for the first time and propose a novel cascaded normalization pipeline, which gives a marked improvement over the baseline.

The remainder of this chapter is organized as follows. In Section 4.1, we give details of the baseline feature set provided by the organizers. In Section 4.2 the experimental setting and results of the Conflict Corpus and in Section 4.3 the application to physical load recognition are given. In both of these sections, the proposed FS methods are applied on the baseline feature set, contrasting the performance of the proposed method with other FS methods and the full feature set. In Section 4.4, the proposed FV based method and the experimental validation on the Eating Condition Sub-Challenge of INTERSPEECH 2015 are given. Each section provides an overview of the methods proposed and concluding remarks on findings. Finally, in Section 4.5, we briefly summarize the work done on other paralinguistic challenge corpora, namely for laughter detection and continuous personality trait mapping.

4.1. Baseline Acoustic Feature Set

The baseline feature set is the same for all three ComParE Challenge corpora experimented in this chapter. This feature configuration is used since INTERSPEECH 2013 [4]. The feature set contains 6 373 features extracted via openSMILE [37] using 54 statistical functionals (c. f. Table 4.2) operating on 65 low-level descriptors (LLD). LLDs cover a wide range of popular Spectral, Cepstral energy related and voicing related descriptors (c. f. Table 4.1).

4.2. Acoustic Conflict Recognition

As mentioned earlier is Section 2.1.2, the state-of-the-art pipeline of paralinguistic speech analysis employs brute-force feature extraction, and the features need to be tailored according to the relevant task. In this work, we extend a recent discriminative projection based feature selection method, namely SLCCA-Filter [9], using the power

Table 4.1. 65 low-level descriptors used in INTERSPEECH 2013 ComParE Challenge.

4 energy related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
55 Spectral LLD
RASTA-style auditory spectrum, bands 1-26 (0–8 kHz)
MFCC 1–14
Spectral energy 250–650 Hz, 1 k–4 kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90
Spectral Flux, Centroid, Entropy
Skewness, Kurtosis, Variance, Slope
Psychoacoustic Sharpness, Harmonicity
6 voicing related LLD
F_0 by SHS + Viterbi smoothing, Probability of voicing
logarithmic HNR, Jitter (local, delta), Shimmer (local)

Table 4.2. Applied Functionals. [1]: Arithmetic mean of LLD / positive Δ LLD, [2]: Only applied to voice related LLD, [3]: Not applied to voice related LLD except F_0, [4]:Only applied to F_0.

Functionals applied to LLD / Δ LLD
quartiles 1–3, 3 inter-quartile ranges
1 % percentile (≈min), 99 % percentile (≈ max) position of min / max
percentile range 1 %–99%
arithmetic mean[1], root quadratic mean
contour centroid, flatness
standard deviation, skewness, kurtosis
rel. duration LLD is above / below 25 / 50 / 75 / 90 % range
rel. duration LLD is rising / falling
rel. duration LLD has positive / negative curvature[2]
gain of linear prediction (LP), LP Coefficients 1–5
mean, max, min, std. dev. of segment length[3]
Functionals applied to LLD only
mean of peak distances
standard deviation of peak distances
mean value of peaks
mean value of peaks – arithmetic mean
mean / std.dev. of rising / falling slopes
mean / std.dev. of inter maxima distances
amplitude mean of maxima / minima
amplitude range of maxima
linear regression slope, offset, quadratic error
quadratic regression a, b, offset, quadratic error
percentage of non-zero frames [4]

of stochasticity to overcome local minima and to reduce the computational complexity. The proposed approach assigns weights both to groups and to features individually in many randomly selected contexts and then combines them for a final ranking. The efficacy of the proposed method is shown in a recent paralinguistic challenge corpus to detect level of conflict in dyadic and group conversations. We advance the state-of-the-art in this corpus using the INTERSPEECH 2013 Challenge protocol [4].

In state-of-the-art computational paralinguistics systems, feature extraction is not a bottleneck, since brute-forcing yields an over-complete feature set. However, this approach requires an elaborate task-relevant pruning of features. This issue comprises the research problem of this section.

This study extends a recent work that uses CCA[1] as a ranking feature selector [9]. It has been used for a variety purposes ranging from multi-view feature extraction [24, 100], to feature selection [9, 10] and regression [19]. Motivated by the success of [9] as well as its limitations, we propose the use of stochasticity to extend [9] by applying CCA between a random subset of features and then by aggregating the feature importance weighted with the canonical correlation value (feature group saliency). The approach is validated on a recent challenge corpus: INTERSPEECH 2013 Conflict Sub-Challenge, where we advance the state-of-the-art using only the audio modality. The proposed CCA based feature filter is introduced in Section 3.1.2, and the work presented here is published in IEEE Signal Processing Letters [11].

The remainder of this section is organized as follows. In the next subsection, a brief review of relevant literature is given. Section 4.2.2 details the challenge corpus. In Section 4.2.3 experimental results are presented. Finally, Section 4.2.4 concludes the work on acoustic conflict recognition.

[1]Background on CCA can be found in Section 2.2.1.1.

4.2.1. Literature Review

Since the primary focus of this chapter is computational paralinguistics, we provide a brief summary of recent paralinguistic works that utilize feature selection in Table 4.3. These papers use different feature selection (FS) methods, like Correlation based Feature Selection [121, 172] and Automatic Relevance Determination (ARD) [173]. While some utilize existing methods, majority of the works listed in the table propose new feature selection methods that better target the specific domain.

Of particular relevance is Random Subset Feature Selection (RSFS), a method that is introduced in [79]. At each iteration, the algorithm selects a random feature set and then measures relevance of each feature based on the performance of the subset that the feature participates in. To compute the relevance, the authors increase the weights of features participating in a set providing higher than average performance by a predefined value p, and similarly reduce the weight by the same amount for the features performing lower than the average. Despite its success, feature and group level weighting are not handled very well, and the method is not scalable as it relies on thousands of simulations.

4.2.2. INTERSPEECH 2013 Conflict Corpus

The INTERSPEECH 2013 Conflict Sub-Challenge [4] aims at automatically analyzing group discussions with the purpose of recognizing conflict. The subject is important since it involves dyadic speech and speaker group analysis in realistic everyday communication. The Conflict Sub-Challenge uses the "SSPNet Conflict Corpus" [181]. It contains political debates televised in Switzerland[2] . The statistics of the corpus are summarized in Table 4.4.

The clips have been annotated following the process illustrated in [182] with respect to conflict level by roughly 550 assessors recruited via Amazon Mechanical

[2]The clips are in French. The data are publicly available and can be accessed from `http://sspnet.eu/2013/09/sspnet-conflict-corpus/`

Table 4.3. Summary of related work in computational paralinguistics employing/proposing feature selection (FS) methods

Work	**Paralinguistic Task**	**Method**
Torres *et al.* (2006) [174]	Depression	GP-Based Two-Stage FS
Park *et al.* (2006) [175]	Emotion	Interactive FS
Torres *et al.* (2007) [176]	Depression	GA-Based FS
Espinosa *et al.* (2011) [177]	Emotion	Bilingual Acoustic FS
Giannoulis and Potamianos (2012) [178]	Emotion	mRMR + SBE
Räsänen *et al.* (2013) [79]	Autism, Emotion and Conflict	Random Subset FS
Kirchhoff *et al.* (2013) [179]	Autism	Submodular FS
Moore *et al.* (2014) [172]	Emotion	Correlation FS
Kaya *et al.* (2014) [9]	Depression	CCA based FS
Kaya *et al.* (2014) [10]	Physical Load	CCA based Multi-view FS
Bejani *et al.* (2014) [180]	Emotion	ANOVA based FS
Kim *et al.* (2014) [173]	Conflict	Automatic Relevance Determination

Table 4.4. Statistics of the Conflict Corpus.

Property	Statistic
# of Clips	1430
# of Subjects	138
# of Females	23 (1 moderator , 22 participants)
# of Males	133 (3 moderator, 120 participants)
# of Political Debates	45
Mean Clip Duration	30 seconds
Conflict Score Range	(−10,+10)

Turk. Each clip is assigned a continuous conflict score in the range [10, +10], giving rise to a regression task. For the challenge, a binary classification task is created based on these labels, namely to classify into 'high' (> 0) or 'low' (< 0) level of conflict. The distribution of instances among partitions (the challenge protocol) is given in Table 4.5.

Table 4.5. Partitioning of the SSPNet Conflict Corpus into train, dev(elopment), and test sets for binary classification [4].

#	Train	Dev	Test	Total
Low	471	127	226	824
High	322	113	171	606
Total	793	240	397	1430

4.2.3. Experimental Results

For the classification task in the Conflict Corpus, we use Support Vector Machines with Linear Kernel, to maximize comparability with previous work on the same corpus. We use Random Forests (RF) to provide an independent classifier benchmark. RF is a combination of decision tree predictors, where each tree is grown with a random (sampled with replacement) set of N instances and a random subset of features [120]. RFs are known to generalize well and are successfully employed in high dimensional pattern recognition. We train SVM models with Platt's Sequential Minimal Optimization (SMO) algorithm [183]. We choose the SVM complexity parameter $\in 10^{\{-5,-4,-3,\ldots,2\}}$.

For RF simulations, we use {10, 20, 30} trees each with a random feature dimensionality sampled in the range of [50, 1000] with steps of 50. For reproducibility, we set the seed of random number generator to one before simulations.

We use the WEKA [184] implementation of Correlation based Feature Selection (CFS) [121] with "Best First" search and SLCCA-Filter methods as independent benchmarks for the Conflict Challenge. We employ Unweighted Average Recall (UAR), which is the mean of individual recalls, as primary evaluation measure:

$$UAR = \frac{1}{K}\sum_{k=1}^{K} TP(k)/P(k), \tag{4.1}$$

where K is the number of classes; $TP(k)$ and $P(k)$ denote the number of true positive instances and total positive instances for class k, respectively. We carry out classification on selected features ranked using both continuous and discretized labels.

Figure 4.1 summarizes the experiments on the training and development sets. The figure shows UAR performances of discretized (class) labels based versus continuous (regression) labels based ranking using SLCCA-Rand and SLCCA-Filter methods in relation to two other baselines. On the overall, we see that the best results are obtained with SLCCA-Rand (blue solid lines) using continuous labels. In the same vein, classification with features ranked by continuous labels are observed to provide better UAR scores than ranking by class labels. We observe that using continuous labels gives a smoother performance contour, improving feature selection for the test set. Moreover, in both types of labels, SLCCA-Rand achieves better performance than other benchmarks.

Finally, we evaluate the proposed method on the challenge test set using the setting that gives the best development set UAR performance. We restrict our test set trials to four: we use the first 500 features that yield the best development set results learned from the training set and the same number of features ranked by training and development sets, together with the two best SVM complexity parameters (0.01 and

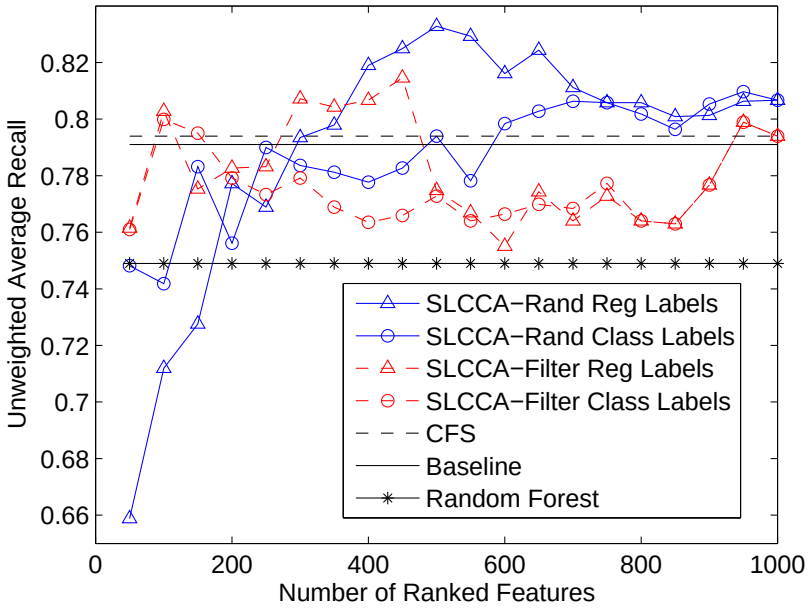

Figure 4.1. Comparison of feature ranking learned from regression and classification labels in relation to challenge baseline [4] and best performance of CFS [121] as an independent feature selector and Random Forest [120] as an independent classifier.

0.001). Using the features learned from the training set, a UAR test set performance of 83.2% is reached. The UAR results improve to 84.6% when the proposed filter method is applied to the combined (training and development) set. The results achieved advance the state-of-the-art UAR (c. f. Table 4.6) on this corpus [79], without resorting to thousands of classification iterations used in [79].

When we analyze the distribution of SLCCA-Rand features yielding the best test set performance with respect to LLD categories, we observe higher proportion of energy- and voicing-related features among the top ranks, compared to MFCC features (c. f. Figure 4.2).

4.2.4. Overview

In this work, we proposed a novel feature selection approach that extends a recently introduced discriminative projection based filter. The proposed approach uses

Table 4.6. Comparison of the highest test set UAR performances using Conflict Corpus with IS 2013 challenge protocol.

Work	Best UAR (%)
Challenge baseline [4]	80.8
Grèzes *et al.* [80]	83.1
Räsänen *et al.* [79]	83.9
Proposed method	84.6
Random Forest [120]	78.6

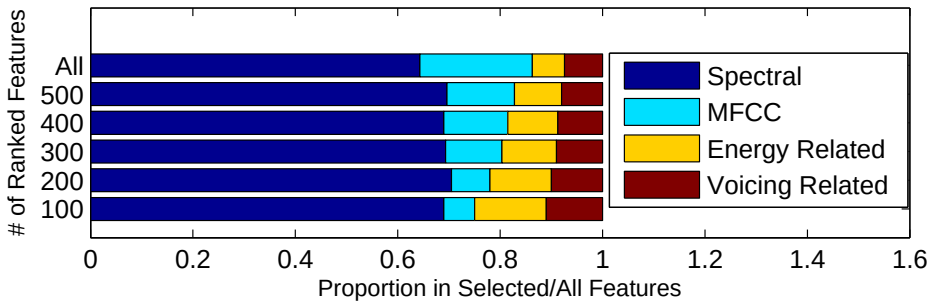

Figure 4.2. Distribution of LLD categories w. r. t. number of ranked features.

the power of stochasticity to overcome the curse of dimensionality by learning feature level and feature group level weights in a variety of random contexts. To maximize the comparison we use the baseline acoustic feature set with SVM Linear Kernel. Ranking features with the proposed method, we advance the state-of-the-art on the Conflict Corpus using the INTERSPEECH 2013 challenge protocol. We observe that learning ranking using regression labels provides better results than using class labels both in SLCCA-Filter and in SLCCA-Rand. The decrease observed in performance with feature selection using class labels is attributed to loss of information during discretization. With regression labels, the continuity in feature space is better mapped to the continuous target variable. Utilizing other methods giving discriminative projections (such as SVM discriminant) and extension to kernel methods constitute our future works.

4.3. Acoustic Physical Load Recognition

In this section, we present our system for INTERSPEECH 2014 Computational Paralinguistics Challenge (ComParE 2014), Physical Load Sub-challenge (PLS). Our contribution is twofold. First, we propose using Low Level Descriptor (LLD) information as hints, so as to partition the feature space into meaningful subsets called *views*. We also show the virtue of commonly employed feature projections, such as Canonical Correlation Analysis (CCA) and Local Fisher Discriminant Analysis (LFDA) as ranking feature selectors. Results indicate the superiority of multi-view feature reduction approach to its single-view counterpart. Moreover, the discriminative projection matrices are observed to provide valuable information for feature selection, which generalize better than the projection itself. In our experiments, we reached 75.35% Unweighted Average Recall (UAR) on PLS test set, using CCA based multi-view feature selection. The paper presenting this work was the winner of this sub-challenge [10].

Prediction of physical load is of interest with many applications [185]. Particularly, classification of the level of heart rate can be benefited in telemonitoring of the disabled/elderly, and in human behavior understanding as it correlates with affective arousal.

INTERSPEECH 2014 baseline feature set contains 6 373 acoustic features [4, 6]. In machine learning literature, utilizing such a high number of features with a small amount of samples is known to reduce generalization power of the learner due to the *curse of dimensionality*. In order to overcome this problem, several feature reduction methods have been proposed in the literature [7, 28]. Fisher Discriminant Analysis (FDA) [97] and Canonical Correlation Analysis (CCA) [94] are two of the most commonly employed statistical methods. While FDA is used in classification problems to project the original features onto a discriminative lower dimensional space, the unsupervised CCA aims at maximizing the mutual correlation of two representations of the semantic object in the respective projected spaces [95].

In a recent study [9], CCA is employed as an acoustic feature selector for contin-

uous depression level prediction. There, the features are exposed to CCA against the continuous labels and are then ranked with respect to the absolute value of their weights in the projection vector (the eigenvector). In the case of classification, this setting is shown to reduce to FDA [98]. One problem in FDA is that it inherently assumes the classes to be unimodal. When classes are composed of several clusters, which is typical in acoustic speech processing, the within class scatter fails to reflect the structure and the projection does not generalize well for pattern recognition. To remedy this problem Sugiyama [105] proposed the incorporation of class-wise neighborhood information as in the Locality Preserving Projection (LPP) in a method called Local FDA (LFDA). LFDA considers the locality of instances of the same class, therefore aims at keeping the structural information in the discriminative embedding. We employ LFDA as an alternative to CCA for feature reduction.

In order to further exploit the data, it is possible to make use of domain knowledge as hints. Some high dimensional data have natural feature partitions that are called *views* in the literature. As these views may be obtained from different modalities, it is also possible to divide the single modality feature set into views. This approach aims at bringing together the self sufficient subset of features to exploit the internal correlations while avoiding the curse of dimensionality. When the multi-view approach is used in filter feature selection, it also helps side-step the *irrelevant redundancy* (IR). IR is incurred when a potential feature that has unique information about the target is omitted due to a high dependency (e. g. correlation, mutual information) with already selected set of features [101]. Moreover, multi-view feature selection reduces space and time complexity, enables processing smaller chunks of data in parallel.

In this section, we propose the use of LLD information of features to partition the massive feature set. From acoustic speech processing, it is known that not all functionals work the same for every LLD. Therefore, our proposed system amounts to selection of functionals per LLD. We also compare and contrast view-based feature selection and extraction using LFDA and CCA. As expected, we obtain better performance (UAR) using multi-view feature selection. We also observe that features selected by means of discriminative projections generalize to unseen data better than

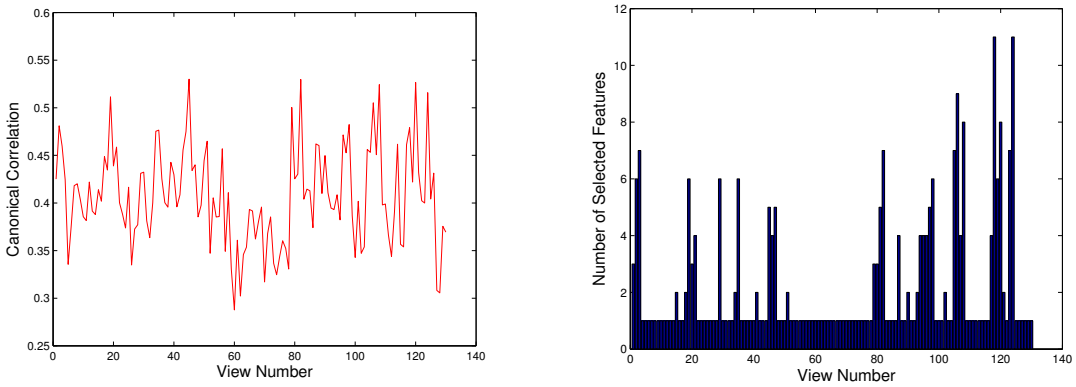

Figure 4.3. Left: Canonical correlation of LLD based views against the physical load labels (Low/High). Right: the number of selected features with CFS per view. The views are sorted lexicographically with respect to the LLD names. The Pearson's correlation between the data of left and right plots is found as 0.65.

the projections themselves.

The layout of this section is as follows. In the next subsection we provide the background on LFDA, then in Section 4.3.1 we introduce the feature selection/extraction scheme. The experimental results are given in Section 4.3.2. Finally Section 4.3.3 concludes with future directions.

4.3.1. Proposed Feature Reduction System

This study is inspired from the discriminative projection based feature selection idea of SLCCA-Filter presented in Section 3.1.1. The study introduced here extends SLCCA-Filter by using LFDA instead of CCA as discriminative projection and applies this base filter method to groups (subsets) of features to improve both computational and classification performance.

While it is possible to obtain features statistically or randomly, in this section, we used the 65 LLDs along with their first derivatives to form the 130 views. Details of the corpus can be found in [6, 186]. The canonical correlations obtained from applying CCA between the LLD based views and the target labels are shown in Figure 4.3. The canonical correlation values range from 0.29 to 0.53.

4.3.2. Experimental Results

In our experiments we utilized the CCA implementation in MATLAB, author's own implementation for LFDA [105]. We used Weka Data Mining tool [184] in classification with Support Vector Machines, and also in our preliminary studies with CFS [121].

4.3.2.1. System Development. To show the superiority of multi-view (MV) versus single view (i. e. full feature set) feature selection independently from the proposed CCA and LFDA based Filter, we used CFS.

In our preliminary experiments, we used a single view feature selection and compared its performance with the features selected from LLD based views. In WEKA CFS implementation, we used BestFirst Forward search option with a backtracking limit of 5 steps. In single view setting the algorithm found 75 features, while in multi view setting 283 features were attained. The distribution of the number of selected features to views is given in Figure 4.3. When the algorithm does not find a good merit in any subset, it generally outputs a single feature.

For classification we used SVMs with Linear and RBF kernels. The precision parameter γ in RBF kernel is set to 0.0005 using cross validation on development set. In both kernels, min-max normalization (min-max Norm) and z-normalization (z-Norm) were tested in the preliminary studies. The set of SVM complexity parameter ranged roughly in double increments, for Linear Kernel we used $\{0.0001, 0.0002, 0.0005, ..., 1\}$ and for RBF kernel a set in the range $0.1 - 50$ is used. The best single view CFS performance is obtained as 61.7 UAR, using z-normalization with RBF kernel with $\tau = 2$. The best multi view CFS performance is given in Table 4.7. As can be seen from the table, all MV settings provide better results than single view setting, z-Norm with RBF kernel yielding relatively better. While these UAR results are lower than baseline development set performance, they motivate further work in MV approach.

Table 4.7. Best SVM performance with multi view CFS features.

Norm	Kernel	τ	UAR (%)
Min-max	Linear	0.02	64.6
Min-max	RBF	50	64.6
Z-norm	Linear	0.001	65.6
Z-norm	RBF	1	66.2

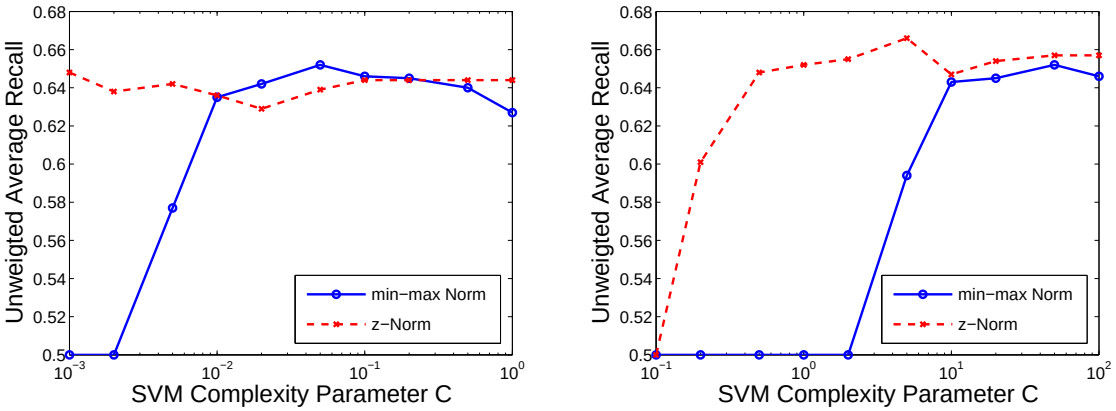

Figure 4.4. Performance of SVM with Linear (left) and RBF Kernel (right) on multi-view LFDA selected features.

4.3.2.2. CCA and LFDA for Feature Selection. Since SLCCA and LFDA provide discriminative projections, which are popularly used in pattern recognition, we also compared the performance of feature transformation against feature selection using the same projection matrices. Encouraged from the preliminary experiments, we use a multi view setting to obtain 130 discriminative features, one from each LLD based view. The best performance (z-Norm, RBF kernel, τ=10) on LFDA projected features was found as 62.9%, the best SLCCA counterpart obtained was 63.5 (z-Norm, RBF kernel, τ=0.2).

Next we issued another set of experiments using 5 highest ranking features from LLD based views with LFDA (5x130=650 features in total). The aim of this set of experiments was to choose an appropriate kernel and normalization method. Similar to preliminary experiments with CFS, we observed that z-norm with RBF kernel worked better than other combinations. See Figure 4.4 for the effect of normalization on Linear and RBF kernels. We finally chose z-Norm along with RBF kernel, where the

τ parameter ranged from 0.1 to 50 with roughly double increments and $\gamma = 0.0005$ as stated before. The number of selected features per view ranged from 10 to 30 with a step of 5 (max features per view is 54). The best results of the two methods are given in Table 4.8. The best overall results are achieved using SLCCA-Filter with 15 features per view (τ=0.1). We observe that most of the results are either better or on par with the challenge baseline on development set (67.2% UAR).

Table 4.8. RBF SVM performance using multi view SLCCA-Filter and LFDA-Filter.

Method	SLCCA		LFDA	
#Feats	τ	UAR(%)	τ	UAR (%)
10	2	67.1	1	67.7
15	0.1	69.4	1	68.4
20	0.1	68.5	1	68.8
25	2	68.9	1	68.0
30	2	66.2	2	67.0

Optimizing the hyper-parameters on the training set, we reached a UAR of 74.16%, beating the test set baseline 71.9%. We further ranked groups of multi-view selected features using mRMR-CCA from [9]. On development set, we obtained the best performance using 1845 features of first 123 views. These relatively slightly refined set of features increased the UAR performance to 75.35% on challenge test set.

4.3.3. Overview

In this section, we examined the filter capability of the discriminative projections specifically Local Fisher Discriminant Analysis and Samples Versus Labels Canonical Correlation Analysis. We also proposed the use of domain knowledge to partition the acoustic feature set into views so as to divide and conquer the data. Using a multi-view approach in feature selection helps avoid the so called irrelevant redundancy, hence allows higher generalization. We show that multi-view setting provides superior scores against its single-view counterpart. Moreover, utilizing the projection matrices for feature selection is found to generalize better to unseen data than the projection

itself. Combining the multi-view approach and the proposed feature selection method we obtain 75.35% UAR on the Physical Sub-challenge.

4.4. Eating Condition Recognition

In this chapter, we address the variability compensation issue by proposing a novel method composed of (i) Fisher vector encoding of Low Level Descriptors (LLD) extracted from the signal, (ii) speaker z-normalization applied after speaker clustering (iii) non-linear normalization of features and (iv) classification based on Kernel Extreme Learning Machines and Partial Least Squares regression. For experimental validation, we apply the proposed method on INTERSPEECH 2015 Computational Paralinguistics Challenge (ComParE 2015), Eating Condition sub-challenge, which is a seven-class classification task. In our preliminary experiments, the proposed method achieves an UAR score of 83.1%, outperforming the challenge test set baseline UAR (65.9%) by a large margin. This work is submitted to participate in the relevant challenge of INTERSPEECH 2015 [22].

The field of paralinguistics has been growing fast in the last decade. A set of paralinguistic tasks, such as emotion [4, 65], depression [16] and personality [64] are popularly investigated, however there is a plethora of other tasks to be discovered.

In this context, INTERSPEECH 2015 ComParE challenge [169] introduces a novel problem, which is to classify the eating condition (EC) of the speaker. There are seven different ECs (speech with no food plus six different types of food) to be classified using acoustic features. The challenge opens a new area of paralinguistic research that can be beneficial for existing studies e. g. by adapting speech and speaker recognizer systems to ECs. The problem is related to a "state" of the speaker, rather than a "trait", and therefore, individual differences should be minimized/compensated.

Modeling/compensating variability due to speakers is of interest in many speech related disciplines. In speaker recognition, state-of-the-art systems are built using i-Vector (i. e. total variability) modeling introduced by Dehak *et al.* [187], which has

its roots in Joint Factor Analysis (JFA) approach [188]. In this approach, the total variance is factorized, and it is postulated that some factors encode for idiosyncratic variations, whereas others are more general. i-Vectors are also used in other paralinguistic tasks [87, 89] for compensating variability due to speakers, rather than augmenting speaker related information for identification purposes.

We propose the use of Fisher vectors (FV) for encoding the low level descriptors (LLD) over utterances. This super vector modeling is introduced and popularly used in computer vision, especially in large scale image retrieval [20, 21]. The idea is to measure the amount of change induced by the utterance/video descriptors on a background probability model, which is typically a Gaussian Mixture Model (GMM). In other words, FV encodes the amount of change of model parameters to optimally fit the new-coming data. This requires the computation of the Fisher information matrix, which is the derivative of the log likelihood with respect to model parameters (hence the name "Fisher"). The encoding requires far less number of components in a GMM than the Bag of Words (BoW) approach [189].

In order to address the speaker variability issue in the EC sub-challenge by employing FV encoding, we first extract Mel Frequency Cepstral Coefficients (MFCC) and RASTA-style Perceptual Linear Prediction (PLP) Cepstrum to represent the signal properties. We show that the combination of RASTA-PLP and MFCC descriptors improve over their individual performances. Moreover, our experiments have shown that the FV encoding of extracted LLDs reaches the performance improvement obtained by the speaker based z-normalization of baseline feature set extracted via openSMILE tool [167]. The performance of the FV is further improved by applying speaker z-normalization. In order to apply this on the challenge test set, where the speaker labels are missing, we implemented Hierarchical Agglomerative Clustering (HAC), which is commonly used to identify speakers [87, 190, 191].

For modeling, we use Extreme Learning Machines (ELM) [13, 14] and Partial Least Squares (PLS) regression [12] based classifiers, motivated by their fast learning capability and outstanding performance in recent challenges [17, 114, 192].

We explain the effect of each component of our framework separately and in a combined fashion. The remainder of this section is organized as follows. In Section 4.4.1 we introduce the proposed method and give background on its major components. The experimental results are given in Section 4.4.2. Section 4.4.3 concludes with future directions.

4.4.1. Proposed Method

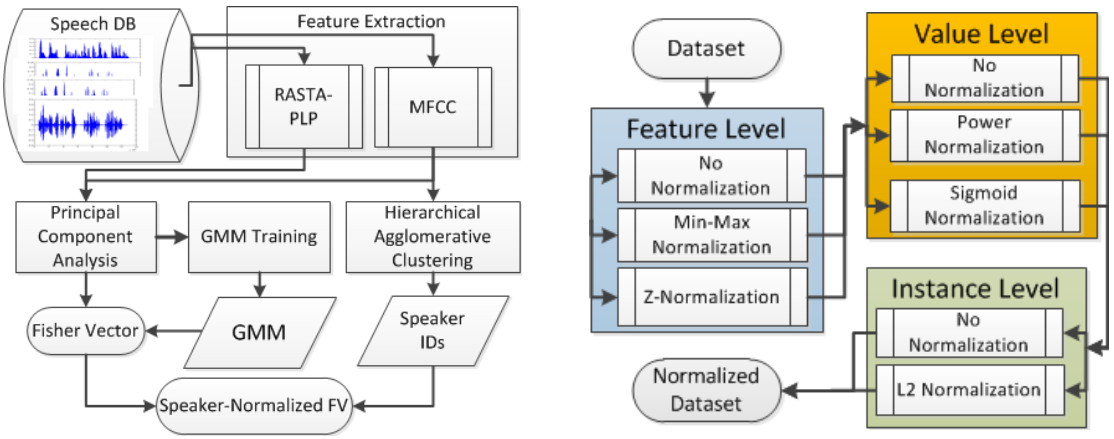

Figure 4.5. Left: overview of the proposed speech signal representation method using Fisher vectors with speaker normalization. Right: Cascaded normalization pipeline.

The overview of the proposed speech signal representation method is given in Figure 4.5, left plot. After this step, non-linear transformation is applied before model learning resulting in a cascaded normalization.

4.4.1.1. Speech Signal Processing. MFCC and RASTA-PLP [29,30] are the most popular descriptors used in a variety of speech technologies ranging from speaker identification to speech recognition, although they are initially designed to minimize the speaker dependent effects. They are also commonly employed in state-of-the-art paralinguistics studies, together with prosodic and voicing related features. Since the task at hand is to recognize the EC, which can be identified with the existence of specific acoustic characteristics (e. g. due to the sound of crunching or chewing), we implement an acoustic model, ignoring the prosody that can be biased towards the speaker identity.

For the purpose of speech signal representation, we extract MFCCs 0-24, and use

a 12th order linear prediction filter giving 13 coefficients. Raw LLDs are augmented with their first and second order delta coefficients, resulting in 75 and 39 features for MFCC and RASTA-PLP, respectively. After a preliminary analysis, we have found that MFCC bands 22-24 are linearly dependent on the first 21 bands; nonetheless, their removal decreased the performance. Moreover, although they are known to be alternative representations, RASTA-PLP and MFCC features are not found to be linearly dependent, therefore they have complementary rather than redundant information.

To distinguish the speech and non-speech frames, we use an energy based voice activity detector. In this approach, frames with lower energy than a threshold τ_E are considered to be non-speech. To smooth the decision boundary, we take the mean energy in a symmetric window with size $2 \times k + 1$, centered at the frame of interest. As a measure of frame-level energy, we tried sum of RASTA-style auditory spectrum and MFCC 0 and observed that thresholding MFCC 0 gives more reliable results on speech signal segmentation.

4.4.1.2. Fisher Vector Encoding. The Fisher vector (FV) provides a supra-frame encoding of the local descriptors, quantifying the gradient of the parameters of the background model with respect to the data. Given a probability model parametrized with θ, the expected Fisher information matrix $F(\theta)$ is the expectation of the second derivative of the log likelihood with respect to θ:

$$F(\theta) = -E[\frac{\partial^2 \log p(\mathcal{X}|\theta)}{\partial \theta^2}]. \tag{4.2}$$

The idea in FV in relation to $F(\theta)$ is taking the derivative of the model parameters and normalizing them with respect to the diagonal of $F(\theta)$ [20]. To make the computation feasible, a closed form approximation to the diagonal of $F(\theta)$ is proposed [20]. As a probability density model $p(\theta)$, GMMs with diagonal covariances are used. A K-component GMM is parametrized as $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ where the parameters correspond to zeroth (mixture proportions), first (means) and second order

(covariances) statistics, respectively. It has been shown that using the zeroth order statistics is equivalent to the BoW model, however in FV, they have a negligible effect on performance [20]. Therefore, only gradients of $\{\mu_k, \Sigma_k\}_{k=1}^{K}$ are used, giving a $2 \times d \times K$ dimensional super vector.

In order to efficiently learn an Acoustic Background Model (ABM) using GMM with diagonal covariances, we first need to decorrelate the data gathered from all utterances. Principal Component Analysis (PCA) is applied on the data for this purpose. To reduce the computational cost, we take LLDs from every second frame to learn PCA and GMM. In our preliminary tests, this sub-sampling did not decrease the performance. Once the parameters of PCA projection and GMM are learned, we use all speech frames from each utterance without sub-sampling to represent them as a FV.

4.4.1.3. Speaker Clustering. Despite the fact that FV encoding aims to compensate the speaker dependent variability, there may still be bias in this representation towards speakers. To further enhance the features in eliminating the speaker dependent information, we use speaker based z-normalization. Since the speaker IDs of utterances are given only for the training/validation set, we need to employ a clustering method to obtain speaker ID information on the challenge test set.

A literature review on speaker clustering reveals that GMM or K-Means clustering on LLDs do not give desirable results. Moreover, the *must-link condition* for LLDs of an utterance is not met with these partitional clustering methods. The most popular method employed for this purpose is single Gaussian based bottom up Hierarchical Agglomerative Clustering with Generalized Likelihood Ratio (GLR) as distance measure [87, 190, 191]. In this method, initialization is done by modeling LLDs of each utterance with a full covariance Gaussian. Then the GLR is computed for each pair of components, and the pair with minimum GLR distance is merged into a single Gaussian component. This continues until one component is left. If the optimal number of components K^* is known, the clustering with K^* components can be taken. Otherwise, one needs to use automatic model selection criteria, such as Bayesian Information

Criterion (BIC) [132], or Minimum Description Length (MDL) [133]. In our problem, the number of speakers in the test set is given in the challenge [169].

In HAC, we use MFCC 1-12 as in [191], instead of 75 dimensional MFCCs used in the ABM. We also use a higher energy threshold compared to the one used in ABM, since here we are interested in clean speech rather than "eating noise" that is useful in discrimination of the EC.

4.4.1.4. Feature Normalization. Perronnin *et al.* further improve the FV representation to be used in linear classifiers (e. g. Linear Kernel Support Vector Machines) with power normalization, followed by instance level L_2 normalization [193]. The authors argue that power normalization helps "unsparsify" the distribution of feature values, thus improving discrimination:

$$f(x) = sign(x)|x|^{\alpha}, \tag{4.3}$$

where $0 \leq \alpha \leq 1$ is a parameter to optimize. In [193] the authors empirically choose $\alpha = 0.5$. In this section, we investigate the suitability of sigmoid function, which is commonly used as hidden layer activation function of Neural Networks:

$$h(\boldsymbol{X}) = \frac{1}{1 + exp\,(-\boldsymbol{X})}. \tag{4.4}$$

This way we avoid a hyper-parameter to optimize, while providing a non-linear normalization into [0,1] range. The flowchart of the normalization steps we applied on the baseline openSMILE features and extracted FVs is given in Figure 4.5, right plot. We use the combination of feature, value (applied to each value of the data matrix separately) and instance level normalization strategies. Without using feature level normalization the performance is poor for the baseline set, while FV encoding does not necessitate this step.

4.4.2. Experimental Results

The corpus of the EC sub-challenge is from [194], which features a total of 1 414 speech utterances produced by 30 speakers with either no food or eating one of the following six food types: apple, nectarine, banana, crisp, biscuit, gummy bear. Each speaker is expected to give data for 7 utterances for each class. However, since some speakers refuse to eat some food, some classes are missing for these speakers. The challenge measure is Unweighted Average Recall (i. e. mean recall of all seven classes), used as challenge measure since INTERSPEECH 2009 challenge [65]. The challenge organizers provide a baseline feature set consisting of 6 373 suprasegmental features extracted via the latest version of openSMILE tool [167].

The data are segmented into a training set with 20 speakers, and a test set with 10 disjoint speakers. Model optimization is done via 20-fold leave-one-speaker-out (LOSO) cross-validation (CV). The test set labels and speaker IDs are not known by the competitors. Further details on the challenge protocol can be found in [169].

For ease of reproducibility, we use open source tools in our experiments. For MFCC and RASTA-PLP feature extraction we use RASTAMAT library [195], for GMM training and FV encoding we use MATLAB API of VLFeat library [196]. PLS regression is a built-in function (plsregress) in the MATLAB version we used in our experiments [197]. Prior to the experiments with FV, we analyze the baseline features with cascaded normalization strategies.

In all our experiments, we generated linear kernels from the preprocessed data and used these kernels in ELM and PLS. The regularization parameter in ELM is optimized in the set $10^{-6,-5,...,5}$ with exponential steps. The number of latent factors for PLS is searched in [2,24] range with steps of two.

4.4.2.1. Experiments with the Baseline Feature Set. As mentioned earlier, the baseline UAR for the training/validation set is computed via LOSO CV by combining the

predictions on each fold to get an overall performance. The baseline UAR scores are 61.3% and 65.9%, for the training/validation and the test sets, respectively.

We first analyzed the features using the combination of normalization strategies described in Section 4.4.1.4. Combination of z-norm + logistic sigmoid + L_2-norm reached the highest LOSO UAR score (63.2%) among other alternatives.

We analyzed the effect of feature selection separately using z-normalization and min-max normalization with two feature filters: multi-view discriminative projection based feature selection [10] that generated the best performance in INTERSPEECH 2014 Physical challenge [6] and a randomized version of this filter [11]. To our surprise, the highest improvement over the baseline was less than one percent, remaining below the individual contribution of the cascaded normalization.

Using the ground truth speaker IDs for speaker z-normalization, it was possible to dramatically increase the UAR performance to 70.1% with PLS and to 70.8% with ELM. When speaker z-norm is augmented with logistic sigmoid + L_2-norm combination, UAR reaches 71.6% with ELM. The results indicate that the features are highly biased towards speakers and a marked improvement can be obtained by minimizing speaker variability.

4.4.2.2. Experiments with the Proposed Method. Considering the effect of the speaker variability, we implemented HAC based speaker clustering for speaker z-normalization. To observe whether this clustering is biased towards the speakers or the classes, we measured the Normalized Information Distance (NID) of clustering to ground truth labels at each iteration. NID is a robust information theoretic measure to compare clusterings suggested by Vinh *et al.* [158]. When two clusterings are identical, the measure becomes zero and when they are totally independent, it takes the value of one. As can be seen in Figure 4.6 (left), HAC is indeed finding clusters of speakers, and the minimum NID (0.04) is found with 21 clusters as opposed to 20 speakers in the training set.

We test the effect of RASTA-PLP and MFCC separately to evaluate their individual and combined performance. We then apply speaker z-normalization using ground truth and predicted speaker IDs for our final system.

Fisher vectors for RASTA-PLP are tested with a range of PCA dimensions, and $K = \{64, 128, 256\}$ components for GMM. The best UAR performance (62.7%) is obtained with 30 PCA dimensions, 128 GMM components, preprocessed using power-normalization + L_2-norm and PLS based classifier. Note that this performance is slightly higher than the baseline.

In the remaining experiments, we use $K = 128$ to train GMMs, as it gives a good compromise between computational complexity and UAR performance. Using MFCC features, the performance is increased to 66.9% with 70 PCA dimensions, with logistic sigmoid + L_2 norm. In results not reported here, we observed that the classifiers react differently to non-linear preprocessing alternatives. We also noticed a jump in UAR performance (64.3%→66.9%) from 60 to 70 PCA dimensions. This implies that the "devil is in the details", as the variability due to eating noise that is useful for discrimination might be contained in these eigenvectors.

When the two descriptors are combined, the best overall UAR is obtained as 70.4%, with 110 PCA dimensions and power-norm + L_2 norm combination. Further dramatic improvement is attained when speaker z-normalization is applied. We reach 76.1% and 77.0% UAR using speaker clustering and real speaker IDs, respectively (see Table 4.9). Score fusion of the best two systems gave 77.5% UAR both with predicted and ground truth speaker IDs.

For the challenge test set, we have submitted predictions of three systems. The first is score fusion of the best two systems with predicted speaker IDs (see the last row of Table 4.9). This resulted in a test set UAR of 81.4%. For the second submission, we re-trained GMM based ABM using the descriptors from the training and test sets. This increased the best training set UAR to 78.9% using logistic sigmoid + L_2-norm combination with PLS. The test set UAR increased slightly to 81.6%. This result is

Table 4.9. UAR scores of RASTA-PLP + MFCC combination.

UAR (%)	Power-L_2		Logsig-L_2		No-norm	
Preprocess	PLS	ELM	PLS	ELM	PLS	ELM
PCA 80	**66.1**	64.2	65.5	64.8	65.7	64.0
PCA 100	**67.5**	63.9	65.7	**67.4**	66.8	65.8
PCA 110	69.4	**70.4**	66.8	67.9	66.0	68.7
PCA 110 with speaker z-normalization						
Real ID	75.3	75.6	**77.0**	**77.0**	76.3	76.5
Pred. ID	74.2	73.9	75.5	74.4	**76.1**	74.1

motivating as it shows that using only training set for acoustic background modeling generalizes as good as combination of the training and test sets. We observed that the predicted labels for the first two submissions differ in 72 instances, therefore we fused their scores for the third submission. This combination reached a test set UAR of 83.1%. The corresponding confusion matrix is given Figure 4.6, where we see a perfect recall of "No Food" class. We also observe high recall for "Crisp" and "Biscuit" classes, where we may expect high confusion. The lowest recall is observed with "Nectarine", which is confused generally with "Apple" (25%) and "Banana" (13%).

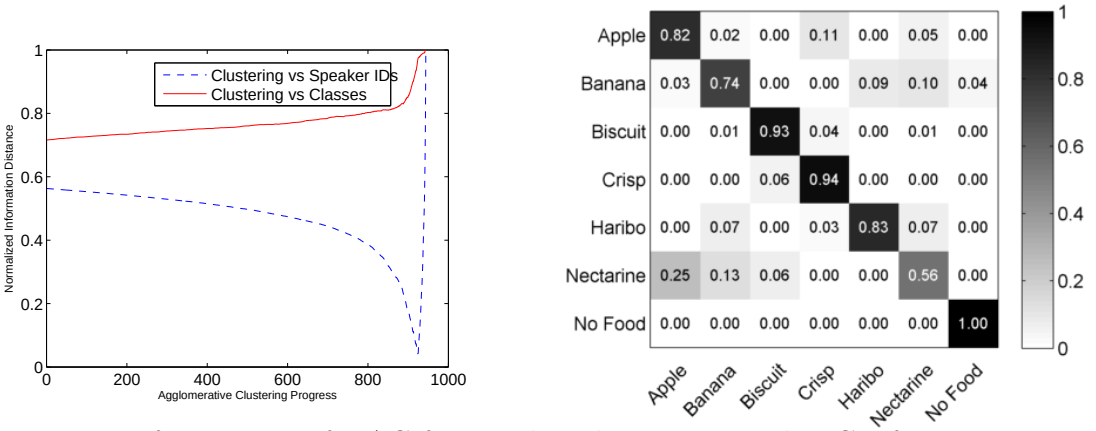

Figure 4.6. Left:Progress of HAC for speaker clustering. Right: Confusion matrix of submission 3 (UAR 83.1%).

4.4.3. Overview

In this section, we proposed a novel method combining the FV encoding with cascaded normalization that is composed of speaker z-normalization and non-linear normalization. The results indicate superior performance of the proposed method over the challenge baselines obtained with openSMILE features. The experiments with the baseline feature set reveal the importance of compensating speaker variability, which is handled partly by the FV approach and partly by speaker z-normalization employed after Hierarchical Agglomerative Clustering. The best overall test set performance is obtained with score fusion of systems trained on combination of RASTA-PLP and MFCC descriptors. The results on both baseline and extracted features indicated that proposed sigmoid normalization is a good alternative to power-normalization used to enhance non-linear discrimination capability of linear classifiers. Future works include application of the proposed method on cross-corpus acoustic emotion recognition.

4.5. Other Paralinguistic Applications

In our contribution to ACM ICMI 2014 Mapping Personality Traits Challenge and Workshop, we propose a system that utilizes Extreme Learning Machines (ELM) and Canonical Correlation Analysis (CCA) for modeling acoustic features [26]. Benefiting from the fast learning advantage of ELM, we carry out extensive tests on the data using moderate computational resources. We further investigate the suitability of our proposed SLCCA-Filter [9] approach to prune the acoustic features, as well as mean smoothing of predictions. In our study, Kernel ELM performed better than basic ELM. Although an average (6-fold cross-validation) Pearson's correlation of 0.642 is reached on the training and validation sets, the overall correlation obtained on the sequestered test set is very low. The test set results indicate the difficulty of the problem and the need for more domain inspired mid-level features. Moreover, obtaining better test set results with flipping the sign for some trait dimensions indicate that the feature-target correlations might be twisted in the training and the test set.

In [25], we investigate several methods on the INTERSPEECH 2013 Paralinguis-

tic Challenge - Social Signals Sub-Challenge dataset. The task of this sub-challenge is to detect laughter and fillers per frame. We apply Random Forests with varying number of trees and randomly selected features. We employ minimum Redundancy Maximum Relevance (mRMR) [122] ranking of features. We use SVM with linear kernel to form a relative baseline for comparability to baseline provided in the challenge paper. The results indicate the relative superiority of Random Forests to SVMs in terms of sub-challenge performance measure, namely Unweighted Average Area Under Curve (UAAUC). We also observe that using mRMR based feature selection, it is possible to reduce the number of features to half with negligible loss of performance. Furthermore, the performance loss due to feature reduction is found to be less in Random Forests compared to SVMs. We also make use of neighboring frames to smooth the posteriors. On the overall, we attain an increase of 5.1% (absolute) in UAAUC in challenge test set.

5. APPLICATIONS IN VIDEO BASED/MULTIMODAL AFFECTIVE COMPUTING

Speech acoustics provide strong cues for arousal based affect classification (e.g. for discriminating anger from sadness), however the valence related cues are weak (e.g. for discriminating anger from happiness). To improve the robustness, usually fusion with other modalities are sought. In this chapter, we provide our audio-visual studies to improve the performance of audio-based affective computing systems.

Studies on multimodal affective computing is of interest in different disciplines, beyond paralinguistics. Some examples are human behavior understanding, social signal processing and biomedical engineering [15, 16]. Using non-verbal behavior to monitor patients of long term disorders such as depression and autism may reduce the overall cost of the health-care systems. Furthermore, affective-responsive applications can be used for rehabilitation of such disorders.

In this chapter, we provide applications of our proposed novel feature selection approaches, as well as problem specific methods to tackle the bottlenecks of the processing pipeline. In the next section, we give our methodology and results on a challenging problem, namely multimodal emotion recognition in natural conditions. This section uses Emotion Recognition in the Wild 2014 Challenge corpus [15]. The results on the baseline audio and video features are presented in our contribution to the challenge [17]. The proposed method is improved with new visual features and a weighted fusion scheme [18], where we reach the state-of-the-art performance with much fewer number of sub-systems compared to EmotiW 2014 challenge winner [114].

Sections 5.2 and 5.3 are about prediction of emotions from video and about audio-visual depression severity level prediction, respectively. Both of problems are included in the AVEC challenges. In Section 5.2, we provide methodology and experiments on AVEC 2014, Affect Sub-Challenge (ASC), where we use only the video modality

for predicting continuous emotion [19]. Section 5.3 reports audio-only, video only and audio-visual fusion results on AVEC 2013 and 2014 Depression Sub-Challenge (DSC), respectively. The methods employed include novel CCA based feature selection methods for audio [9] and CCA based depression covariate extraction for video [24].

In addition to audio-visual fusion, we have optimized the use of individual modalities. Particularly for the visual modality, face images were processed to estimate expressions and depression levels. True to the spirit of the entire work, we sought a small set of descriptive and discriminative features in the visual modality. By dividing the face image into disjoint subsets (i.e. regions), we learned from the training set which regions were informative for the task at hand. Our findings pointed to the inner facial regions specifically, and we obtained improvements over using the entire face. As a general observation of the "less is more" principle, using a small number of well-selected features is the better approach in these high dimensional tasks.

5.1. Multimodal Emotion Recognition in the Wild

Emotion recognition from video and audio is gaining increasing attention, especially because its outputs can be used in many related domains [2]. In the last decade, a considerable amount of research efforts spent in the field was on controlled, laboratory-condition data. In some of such corpora (e. g. Berlin Emotional Speech Database [49]) it was possible to obtain classification scores even better than human perception [1]. Now the field is moving on to less controlled conditions, including noisy audio-visual background, large variance in facial appearance and spoken content.

Audio-visual emotion related challenges have been instrumental in improving the state-of-the-art in this field. The challenges provide a great opportunity for the researchers in the field and help advance the state-of-the-art by bringing together experts from different disciplines, such as signal processing and psychology. One such challenge series is Emotion Recognition in The Wild (EmotiW) [15, 198] that provides out of laboratory data -Acted Facial Expression Wild (AFEW)- collected from videos that mimic real life [42].

In this study, we apply a powerful classification paradigm, Extreme Learning Machines (ELM) to audio-visual emotion recognition. We investigate feature/group selection in both modalities to enhance generalization of learned models. We further extract audio features using the most recent INTERSPEECH Computational Paralinguistic challenge baseline set [4] with the freely available openSMILE tool [37] and augment the AFEW dataset with four other publicly available emotional corpora: Berlin EMODB [49], Danish Emotion Database (DES) [48], eNTERFACE Database [50], and the Turkish Emotional Database (BUEMODB) [23]. We carry out score level fusion of modality-specific models, which boosts the performance of individual models.

Further contributions of this section are as follows. In addition to baseline feature sets, we use new visual feature types and compare ELMs with Partial Least Squares based classifier, which yields the best performance in the state-of-the-art system on the EmotiW 2014 Challenge [114]. We extract dense SIFT features from images, representing the videos (image sets) using SVD based linear subspace, covariance matrix and the Gaussian distribution statistics, all of which lie on Riemannian manifolds. In line with [114], Riemannian kernels are used in classifiers. We also extract video features using Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP), which is shown to be less sensitive to registration errors compared to LGBP and LBP-TOP [111].

The remainder of this section is organized as follows. In the next subsection, we provide background on ELM. Then in Section 5.1.1 we overview the corpora and describe our proposed approach. In Section 5.1.3 we give experimental results, and conclude in Section 5.1.4.

5.1.1. The AFEW Corpus

The EmotiW 2014 Challenge [15] uses the AFEW 4.0 database, which is an extended version of AFEW 3.0 used in the EmotiW 2013 Challenge [198]. The corpus contains videos that are clipped from movies with the guidance of emotion related keywords in the movie script for the visually impaired [42]. The 2014 challenge provides

Figure 5.1. Illustration of aligned images with varying conditions.

a total of 1 368 video clips collected from movies, representing close-to-real-world conditions [42]. The corpus is partitioned into training, development and test sets. The challenge participants are expected to develop their system using the first two sets and send their predictions for the test set, whose labels are sequestered.

The nature of the data collection poses challenging conditions, e. g. in terms of background noise, head pose and illumination. In most of the clips there is a single actor. However, in some cases there are multiple actors on the scene. Speech is generally accompanied by the background music and noise. While the challenge annotations come with a standard face detector result, the difficult conditions cause problems even in the early stages of the processing pipeline. Some example aligned images illustrating this problem are shown in Figure 5.1. We observe that in addition to precisely aligned frontal faces, there are misaligned or occluded faces, or images that do not contain faces.

5.1.1.1. Baseline Feature Sets. The baseline video features consist of Local Binary Patterns from Three Orthogonal Planes (LBP-TOP), compacted via uniform LBP [110] extracted from the detected and aligned faces in the videos. To add structural information to the LBP histogram representation, the face is divided into non overlapping

$4 \times 4 = 16$ regions and an LBP histogram is computed per region. The TOP extension applies the relevant descriptor on XY, XT and YT planes (T represents time) independently and concatenates the resulting histograms. In total, we have $59 \times 3 = 177$ dimensional visual descriptors per region.

The baseline audio features are extracted via freely available openSMILE tool [37] using INTERSPEECH 2010 Paralinguistic challenge baseline set [168]. The 1 582 dimensional feature set covers a range of popular LLDs such as Fundamental Frequency ($F0$), MFCC [0-14], Line Spectral Pairs Frequency (LSP) [0-7] mapped to a fixed length feature vector by means of functionals such as arithmetic mean and extrema.

5.1.2. Exracted Features

In our experiments, we use the aligned faces provided by the challenge organizers for visual signal processing. The images are first resized to 64×64 pixels. In the preprocessing step, we use PCA based data purification as shown to be effective in [199, 200]. The idea is to measure the mean reconstruction error per image $x_i \in \mathbb{R}^D$ with $Err_i = \frac{1}{D}||(x_i - \mu) - W_{pca}^T W_{pca}(x_i - \mu)||$, where $\mu \in \mathbb{R}^D$ is the training set mean vector, and W_{pca} is the reduced PCA projection coefficient matrix. We discard the frames with a high reconstruction error as these are probably poorly detected or aligned images. In our study, we use the L_1 norm and remove the videos that have less than three valid images from training and validation sets. In our preliminary studies on AFEW 4.0, we observe a considerable accuracy increase due to purification.

We first extract dense SIFT features from images, due to their popularity in compact representation of local appearance [201]. The dimensionality of image features are reduced via PCA, whose coefficients are learned from the training set, prior to video modeling. We use the same parameters as in [114] to extract SIFT features: typical 128 dimensional SIFT descriptors are extracted from 16×16 pixel patches with steps of 8 pixels that gives $7 \times 7 = 49$ overlapping blocks. Therefore, the dimensionality of the concatenated SIFT feature vector is $49 \times 128 = 6\,272$. As mentioned earlier, the feature dimensionality is reduced via PCA, preserving 90% of the total variability.

In addition to dense SIFT features based video representation, which will be detailed below, we also implemented Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) feature representation [111]. The basic idea here is to combine the power of static LGBP descriptor and dynamic LBP-TOP. The work of Almaev and Valstar has shown that LGBP-TOP descriptor outperforms LGBP and LBP-TOP for Facial Action Unit (FAU) recognition task and it is less susceptible to rotation errors compared to these methods [111].

Video Representation for Dense SIFT Descriptor. After extraction of image features, the image sets are represented via four alternatives from which kernels are obtained. The first and simplest approach is using statistical functionals to provide a baseline. Here we use mean and range of image features over frames. Let $X_v \in \mathbb{R}^{d \times F_v}$ be the matrix representing d dimensional features of video v having F_v frames. Using mean and range functionals results in a $2 \times d$ dimensional video feature vector. We use three video modeling approaches that represent image sets in Riemannian manifold: SVD representation, Covariance matrix and the Gaussian distribution. The details for these video representations and the extracted kernels can be found in Section 2.2.2.4.

LGBP-TOP. Details of the spatio-temporal LGBP-TOP descriptor used in the thesis are given in Section 2.2.2.3. The parameters of the Gabor filter are defined up to the 2D sinusoid and the 2D Gaussian used in the convolution. Using three spacial frequencies ($\phi = \{\pi/2, \pi/4, \pi/8\}$) and six orientations ($\theta = k\pi/6, k \in \{0..5\}$), we form a set of 18 Gabor filters. The dimensionality of the feature vector is therefore $2 \times 18 \times 16 \times 58 \times 3 = 100,224$.

5.1.3. Experimental Results

In our experiments we test the suitability of basic and kernel ELM for the problem of audio-visual emotion recognition. To probe the individual performance, we handle the video and audio separately and then combine the decisions of best performing

modality-specific ELMs. Fast ELM training gives us the ability to simulate a wide range of hypotheses with moderate system requirements.

5.1.3.1. Experiments with Baseline Visual Features. For visual classification, we assess the contribution of different facial regions. Apart from reduced dimensionality for better modeling, the reason of focusing on a small number of facial regions are (i) partial peripheral occlusion of face, (ii) cluttered background in case the face is tilted and (iii) robustness of different regions to alignment issues.

We issue ELM tests with Linear and RFB kernels on various facial region combinations. For Kernel ELM, we experimented different regularization parameter values $\tau \in 10^{\{-5,\ldots,5\}}$. The best results for each region and kernel type are provided in Table 5.1. We see that using 2-by-2 inner facial regions or 2-by-4 vertical middle regions (midface vertical), we can obtain better performance than the full set of 16 regions. Our results indicate that for emotion related tasks with difficult registration conditions, focusing on the inner face reduces the feature dimensionality, while preserving discriminative information, and results in better classification rates. Note that we carry out group-wise feature selection, since individual selection of histogram features are not meaningful and sufficient for pattern recognition.

To keep the table uncluttered, the parameters yielding the reported results are not included. The best RBF kernel results using at least four regions are obtained with scatter parameter $\sigma = 10^3$ and regularization parameter $\tau = 10$. On the other hand, best results with linear kernel are obtained with $\tau = 10^{-2}$. Both in video and audio modalities, we record the hyper parameters giving the best results to be later used in test set predictions.

5.1.3.2. Experiments with Baseline Acoustic Features. After the first probe into the full set of baseline acoustic features, we applied several feature selection (FS) methods. This was followed by extraction of a more recent and larger feature set that was used in INTERSPEECH 2013 (see Section 2.1.2 for details) via openSMILE tool [37].

Table 5.1. Validation set accuracy comparison of facial regions using Linear and RBF Kernel ELM.

Facial Regions (#)	**RBF**	**Linear**
Whole Face (16)	39.35%	38.81%
Midface Horizontal (8)	37.47%	39.08%
Midface Vertical (8)	38.89%	40.16%
Inner face (4)	39.89%	39.08%

We first used the iterative minimum Redundancy Maximum Relevance (mRMR) filter [122] for FS. In addition to mRMR, we applied Samples versus labels CCA (SLCCA-Filter) to Low Level Descriptor (LLD) based feature groups and then concatenate the ranked k features from each group as in [10]. When all features are subjected to CCA against the labels, the absolute value of the projection matrix $\mathbf{V}$ can be used to rank the features [9]. We extend the LLD based approach using mRMR and combine top $k = \{5, 10, 15, 20\}$ ranking features from each LLD group. For mRMR, the first 200 features are tested with steps of 10, each with the set of ELM hyper parameters discussed in previous sections. Together with regular mRMR, we test three FS approaches on baseline acoustic features.

The best validation set results of FS approaches utilized in the study are given in Table 5.2. We observe that the best LLD-based SLCCA-Filter outperforms the best performance obtained with LLD-based mRMR. However, no FS method performs better than the full set of features. The superior performance of the full set can be attributed to the ELM learning rule, which minimizes the norm of the projection, thus making use of all features without over-fitting. In a regular Neural Network where the weights are learned via gradient descend based back-propagation, FS would help avoid over-fitting, therefore may yield better results than the full set. Another reason that a selected subset does not perform better than the full set is the fact that the paralinguistic information can be distributed over a wide range of features. This is the reason why state-of-the-art results in the field are obtained with very high dimensional supra-segmental acoustic features [1].

Table 5.2. Validation set performance comparison of feature selection methods.

	RBF	**Linear**
All Baseline Feats	35.77%	35.77%
mRMR Ranked Feats	33.16%	34.46%
mRMR-LLD (k=10)	33.68%	32.38%
SLCCA-LLD (k=10)	35.51%	35.25%

We further included four other publicly available emotional corpora to test whether additional corpora would improve training or not. These are Berlin EMODB [49], Danish Emotion DB (DES) [48], eNTERFACE DB [50], and the Turkish Emotional DB (BUEMODB) [23]. Note that all corpora are acted, although two are recorded in studio. We use the instances belonging to the seven classes of AFEW.

Cross corpus evaluation results are given in Table 5.3. Class distribution and some basic information about the corpora are given in Table 5.4. All corpora are individually normalized to range [-1,1]. Without corpus-wise normalization, the corpora are found to impair the generalization of the learner. This finding is in accordance with cross corpus work of Schuller *et al.* [3]. We see a performance decrease with respect to given baseline features. eNTERFACE and EMODB provide some performance increase with respect to INTERSPEECH 2013 baseline set whereas BUEMODB and DES do not contribute at all. We see that even when all additional corpora are included, the performance is below the accuracy obtained via only EmotiW 2014 Challenge features.

Table 5.3. Best validation set performance of multi corpus training.

Corpora	**Accuracy**
AFEW 4.0 IS13 Features	33.42%
+ eNTERFACE	34.20%
+ EMODB	34.20%
+ BUEMODB	33.42%
+ DES	33.42%
+ All corpora	34.20%

Table 5.4. Class distribution of additional emotional corpora. Classes correspond to A(nger), D(isgust), F(ear), H(appiness), N(eutral), SA(dness), SU(rprise). All corpora are acted and the spoken content is the same for all subjects/emotion classes.

Corpus	**Content**	**A**	**D**	**F**	**H**	**N**	**SA**	**SU**	**#All**
EMODB	German	127	38	55	64	78	53	-	415
DES	Danish	85	-	-	86	85	84	79	419
eNTERFACE	English	200	189	189	205	-	195	182	1170
BUEMODB	Turkish	121	-	-	121	121	121	-	484

5.1.3.3. Comparison of ELM with PLS based Classifier. Here, we compare the performance of Kernel ELM with the Partial Least Squares regression based classifier used in [114], which reports state-of-the-art results for EmotiW 2014 Challenge. For the details of PLS regression, the reader is referred to [12]. In [114], PLS regression is applied to classification in one-versus-all setting, then the class giving the highest regression score is taken as prediction.

We compare the two classifiers first on challenge baseline features. The best validation set results of two methods, and corresponding test set results of ELM are given in Table 5.5. Note that the audio only results given here with PLS are higher than those reported in [114]. This is because we use kernels, while in [114] the acoustic features were used directly. Analyzing the scores on the baseline feature sets, we observe an overall better performance with ELM, and the margin increases with modality-fusion. To show that the validation set scores are highly indicative of the test set performance, we give the test set results of ELM on the right most columns of Table 5.5. Note that as discussed earlier, the inner facial regions generalize better than the whole face due to less sensitivity to occlusion and registration errors.

For further comparison on extracted dense SIFT features, we experiment on statistical functional based video representation. Using only mean and range statistics gives 39.84% and 41.19% validation set accuracy for PLS and ELM, respectively. Note that ELM performance here is higher compared to the best video only result on baseline features. This might be partly due to data purification. Finally, we compare the

two methods using the six Riemannian Kernels described in Section 2.2.2.4. The best validation set performances are listed in Table 5.6. In comparative experiments, we use the same kernels for two methods, optimizing their hyper-parameters on the validation set. Note that the PLS performance using dense SIFT is slightly lower compared to those reported in [114], which may be attributed to the number of PCA eigenvectors prior to video representation. Similar to experiments on baseline features, we observe better overall performance with ELM, giving higher than 43% accuracy on Grassmanian kernels (SVD). When we probe the test set performance of the best models (SVD representation with Linear Kernel), we get accuracies of 40.29% for PLS and 43.23% for ELM, respectively. This difference is not found to be statistically significant with McNemar's test [202]. While the results confirm the good performance of PLS as classifier, it is also clear that the state-of-the-art performance of [114] is largely due to using an ensemble of 24 visual systems, which complement each other.

Table 5.5. Comparison of PLS and ELM performance on EmotiW 2014 baseline feature sets. IF: inner face, WF: whole face

Accuracy (%)	**Validation**				**Test**
Kernel	**Linear**		**RBF**		**RBF**
Classifier	PLS	ELM	PLS	ELM	ELM
Video (WF)	39.08	39.89	38.27	39.35	36.11
Video (IF)	37.74	39.08	39.08	39.89	39.07
Audio	35.25	35.77	34.46	35.77	37.84
Fusion (WF)	41.51	43.13	39.62	42.86	43.00
Fusion (IF)	40.16	42.32	40.43	44.20	44.23

Table 5.6. Comparison of validation set accuracies of PLS and ELM on Riemannian Kernels for video representation.

	SVD		**Covariance**		**Gaussian**	
	Linear	RBF	Linear	RBF	Linear	RBF
PLS	41.46	40.92	38.21	40.65	39.84	37.94
ELM	43.63	43.09	39.84	41.46	39.30	39.02

Lastly, we compare the performance of the two classifiers on extracted LGBP-TOP features. We optimize the σ parameter of the Gabor kernel by observing its effect on the Gabor pictures. On the overall, no optimization is done for other parameters of the Gabor filters. Considering the massive dimensionality, filter and feature selection have a high potential of improving generalization. This is left for future work. Inner facial regions in LGBP-TOP did not provide a performance increase as in baseline LBP-TOP features. We attribute this to the added data purification step, which eliminates partially occluded or badly aligned faces. For linear kernels, the best validation set performances are 42.05% and 39.35% for ELM and PLS, respectively. With RBF kernel, the best accuracies are 41.78% for ELM and 41.51% for PLS.

All our results are obtained on powerful feature sets with good preprocessing. Subsequently, while the ELM classifier usually reaches higher accuracies than the PLS classifier, these differences were not significant. We have recently contrasted these classifiers on a new emotional speech corpus (EmoChildRU), which is collected from $3-7$ years old Russian children in natural conditions [165]. The data are annotated for three valence related affective classes: comfort, discomfort and neutral. Our results indicate that PLS is highly sensitive to preprocessing and to feature representation, whereas ELM consistently gives (in most cases significantly) better results.

5.1.3.4. Multimodal Fusion and Test Set Results. We finally test the best performing modality-specific systems using equal-weighted fusion (EF) and weighted fusion (WF) schemes. In EF we average the class-wise predictions to get a fused score, whereas in WF class-wise weights are used for each sub-system. Using ELM with the baseline features, the best performing single modality systems give 37.84% (audio full set) and 39.07% (video innerface) accuracy on the validation set. Using the extracted features from purified images, we observe 42.05% accuracy with LGBP-TOP and 43.63% accuracy with dense SIFT (Linear Grassmanian kernel).

We first analyze EF on the modality-specific ELMs learned on the baseline features. Then we combine the best modality-specific ELMs using WF, where the optimal

fusion weights are searched over a random pool of fusion matrices (similar to the approach taken in [203]). For this we randomly generate 50.000 fusion matrices for each alternative combination, and normalize them over the models. To avoid over-fitting on the validation set, fusion weights are rounded to three decimal digits. Since RBF kernels are observed to give better performance in decision fusion, all combination experiments are carried out with scores obtained from RBF ELMs.

Validation and test set performances of multimodal decision fusion of ELMs learned on baseline feature sets are provided in Table 5.7. We observe that using inner facial regions in audio-visual fusion provides better generalization than the full set of features. We also see that multimodal fusion with midface features provides the highest validation set accuracy, but this does not yield a high score on the test set. This might be attributed to over-fitting to validation set, however the hyper-parameters are not specifically optimized for this system. The most probable reason of the high discrepancy between validation and test set performances of systems using mid face vertical features is partial occlusion and registration errors. Thus, the higher generalization performance of inner facial features is not only due to the relevant information they contain, but also due to resilience to occlusion and environmental noise.

Table 5.7. Validation and test set accuracies (%) for equal weight decision fusion of modality-specific kernel ELMs trained on baseline feature sets.

System	**Val**	**Test**
LBP-TOP (Midface Vertical) and Audio SLCCA-LLD (k=10)	47.17	38.33
LBP-TOP (Midface Vertical) and Audio All	44.47	38.08
LBP-TOP (Inner face) and Audio All	44.20	44.23
LBP-TOP (Inner face) and Audio SLCCA-LLD (k=10)	43.13	43.98
LBP-TOP (Wholeface) and Audio All	42.86	43.00

The test set confusion matrices for systems obtained with audio baseline features, video baseline features and their fusion are given in Figure 5.2. The diagonal elements indicate the recall of the corresponding classes. On the overall, we observe that fusion system boosts the performance of single modality systems. However, since the audio based system does not recognize Disgust, Happiness and Surprise classes well (if at all),

the fusion system shows a lower recall in these classes compared to the video based system. On the other hand, recall performance of the audio based system outperforms the video based system in the remaining four classes. These results imply that a confidence based fusion of modality-specific systems can advance the overall recognition. Therefore, we use a weighted fusion scheme in further experiments, where we employed combinations of three and four sub-systems.

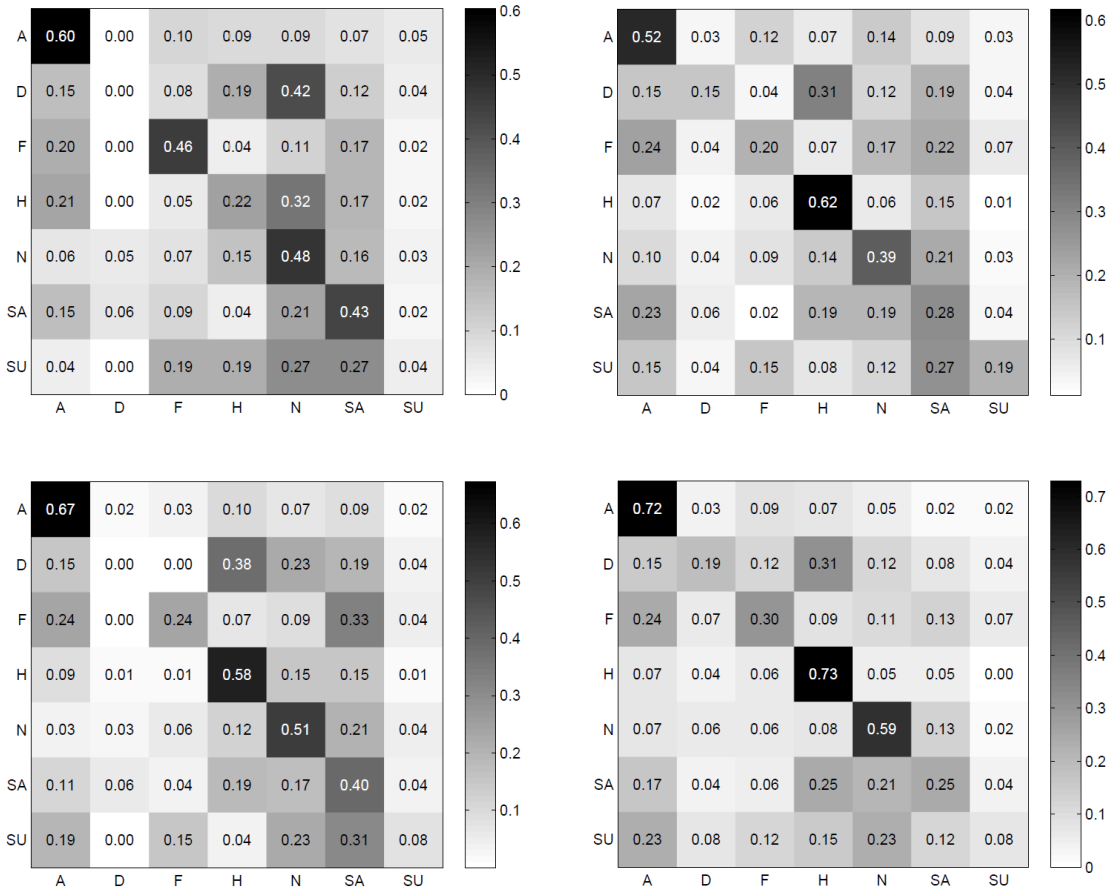

Figure 5.2. Top row: test set confusion matrices for Kernel ELM models trained on baseline audio (left) and video (right) features. Bottom row: test set confusion matrices for multimodal score fusion systems. Left: unweighed fusion of models trained on baseline audio and video features, right: the best performing weighted score fusion system. Classes correspond to A(nger), D(isgust), F(ear), H(appiness), N(eutral), SA(dness), SU(rprise).

Weighted score fusion of LBP-TOP (inner face), SIFT SVD and Audio All gave a validation set accuracy of 49.04% that rendered a test set performance of 46.44%. Inclusion of LGBP-TOP in this scheme led to accuracies of 51.49% and 50.12%, on

the validation and the test set, respectively. It is worthy to note that using a weighted fusion of four sub-systems, we reach the state-of-the-art test set performance obtained by [114] that combines 25 sub-systems. Fusion weights used in the best system and corresponding confusion matrix can be found in Table 5.8 and Figure 5.2, respectively.

Table 5.8. Fusion weights for the best performing system. Classes correspond to A(nger), D(isgust), F(ear), H(appiness), N(eutral), SA(dness), SU(rprise).

	SIFT SVD	LGBP-TOP	LBP-TOP	Audio
A	0.006	0.254	0.377	0.363
D	0.277	0.028	0.656	0.039
F	0.095	0.285	0.097	0.523
H	0.052	0.458	0.238	0.252
N	0.307	0.237	0.131	0.325
SA	0.398	0.303	0.215	0.084
SU	0.446	0.034	0.306	0.214

5.1.4. Conclusions and Outlook

In this section, we introduce ELMs for audio-visual emotion recognition in the wild. ELMs provide accurate results with several orders of magnitude faster training compared to SVMs and SLFNs. Typically, this leads to more time for parameter search and optimization. We test facial feature group selection as well as recently proposed acoustic feature selection approaches for this problem.

We compared ELM with a PLS based classifier that is used in the top system of EmotiW 2014, and obtained better results with ELM. We achieve the best validation and test set results with decision fusion of modality-specific ELM models. While our results verify the importance of multimodal fusion and combination of diverse classifiers, they also highlight the importance of the fusion strategy.

The tested systems performed very poorly on some of the classes. In particular,

it was very difficult to classify happiness and surprise from audio, whereas fear and sadness are difficult to classify from video. Disgust is difficult for both modalities. This result shows that in-the-wild emotions are much more difficult to recognize compared to controlled conditions typically used in the literature.

Our tests with additional speech corpora to augment training did not contribute to accuracy. One possible cause for the lack of improvement is the difference in the acquisition conditions of the corpora. Furthermore, acoustic feature selection was not found to improve the performance. On the other hand, in video modality using a semantically meaningful subset of facial regions, it was possible to obtain better recognition results than the full set, both in the development and the test set.

5.2. Ensemble CCA for Continuous Emotion Prediction from Video

This section presents our work on ACM MM Audio Visual Emotion Corpus 2014 (AVEC 2014) using the baseline features in accordance with the challenge protocol. For prediction, we use Canonical Correlation Analysis (CCA) in affect sub-challenge (ASC). The video baseline provides histograms of Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) features. Based on our preliminary experiments on AVEC 2013 challenge data, we focus on the inner facial regions that correspond to eyes and mouth area. We obtain an ensemble of regional regressors via CCA. We also enrich the 2014 baseline set with Local Phase Quantization (LPQ) features extracted using Intraface toolkit detected/tracked faces. Combining both representations in a CCA ensemble approach, on the challenge test set we reach an average Pearson's Correlation Coefficient (PCC) of 0.3932, outperforming the test set baseline PCC of 0.1966.

The most recent ACM MM Audio-Visual Emotion Corpus and Challenge (AVEC 2014) focuses on prediction of self-reported level of depression and multiple rater human annotated 3 affective dimensions (arousal, valence and dominance) [16]. The organizers provided baseline video and acoustic features along with the video clips partitioned into training, developments and test sets.

In this section, we use an ensemble of Canonical Correlation Analyzers (CCA) to predict affective dimensions from video modality. CCA is a statistical method that finds linear projections for two views (representations) of a semantic object to maximize the mutual correlation [94, 95]. CCA has a variety of applications in pattern recognition ranging from multimodal fusion [95, 100] to feature selection [9, 101]. CCA also has successful applications in affective computing [204, 205]. In [205] Shan *et al.* use CCA for feature level fusion of body gestures and facial expressions on gesture and emotion recognition tasks. In [204] the authors utilize CCA introducing Matrix CCA (MCCA) for facial action recognition and facial parts synthesis.

A previous work on AVEC 2013 ASC [206] applies CCA between MFCC based LLDs and the affective dimensions. Another study on AVEC 2013 benefits from CCA as an acoustic feature selector for depression, where the authors utilize the projection vector to rank the features [9]. Here, we use CCA for feature extraction and regression in video modality. In a recent work on the same corpus, CCA is used for extraction of audio-visual depression covariates [24].

To allow comparability with AVEC 2013 baseline set [5] and to enrich the 2014 baseline video feature set, we extract Local Phase Quantization (LPQ) features. In both feature representations we focus on regions corresponding to eyes (including eyebrows) and mouth area, as these are found to be the most informative in our preliminary experiments with AVEC 2013 corpus. We attribute the better performance of eyes and mouth to information they carry about the affective state, while the remaining regions are thought to show stronger tendency to represent the speaker identity.

The remainder of this section is organized as follows. In Section 5.2.1 we briefly introduce corpus and baseline feature sets. In Section 5.2.2 we give the experimental results. Finally, Section 5.2.3 gives an overview.

5.2.1. The Corpus and Features

AVEC 2013 and 2014 [5, 16] use a subset of the audio-visual depressive language corpus (AVDLC), which includes 340 video clips of subjects performing a Human-Computer Interaction task while being recorded by a webcam and a microphone. In AVDLC, the total number of subjects is 292 and only one person appears per clip, i. e. some subjects feature in more than one clip. The clip duration ranges from 20 to 50 minutes, with a total duration of 240 hours and an average of 25 minutes. The age of subjects ranges from 25 to 63. The target variable range is [-1,1] for the three affective dimensions [5].

Recorded behavior includes speaking out loud while solving a task, counting from one to 10, reading excerpts of a novel and a fable, singing, free talk: telling the best event and a sad event from childhood. The depression levels were labeled per clip using Beck Depression Inventory-II (BDI-II) [207], a subjective self-reported 21 item multiple-choice inventory.

For the AVEC 2014 challenge, only two of the 12 tasks from AVDLC are used. These are referred as Freeform and Northwind tasks. In the Freeform task the subject is asked to recall a good or bad memory from the past using her own words. In the Northwind task a German fable about a competition between Sun and the Northwind is read by the subject. This story has markedly depressing undertones, with the main character facing failure and giving up after experiencing helplessness. For both tasks, the recordings are split into three partitions: training, development, and test sets of 50 recordings each, respectively.

5.2.1.1. Baseline Audio Feature Set. For audio modality, a set of 2 268-length openSMILE [37] features, which were introduced in AVEC 2013 [5], are provided to participants. The acoustic feature sets are arranged in three segmentation settings: short (3 s. overlapping frames with 1 s. shifts), long (20 s. overlapping frames with 1 s. shifts), voice-activity detected (VAD) segments. In VAD segmentation a voice activity

detector [208] is used to split the clip when there is a pause for more than 200 ms. The statistical functionals (e. g. moments, extremes) are then applied on the Low Level Descriptor (LLD) contours (e. g. MFCC 1-16, $F0$, jitter) of segments. The reader is referred to the paper on the challenge [16] for further details of acoustic features. In AVEC 2013 challenge paper, features of short and long segmentations are reported to work well on affect and depression, respectively [5].

5.2.1.2. Video Feature Sets. The baseline video features of AVEC 2014 consist of Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [111]. LGBP-TOP combines the power of Gabor wavelet representation of image with TOP extension (see Section 2.2.2.3 for details). Gabor wavelet is obtained from convolution of a Gaussian and a sinusoid. With varying rotation and phase angles, a set of Gabor pictures are obtained for each video frame then Local Binary Patterns (LBP) are computed for three orthogonal planes (i. e. XY, XZ and YZ). The patterns are finally represented as a histogram. Since this histogram representation does not keep the structural information of facial features, the face is divided into $4 \times 4 = 16$ regions and an LGBP-TOP histogram is computed per region. In AVEC 2014 baseline set, 18 Gabor pictures with 59 dimensional uniform pattern LBP [110] are computed on XY plane resulting in 1062-dimensional histograms per region.

To accompany the baseline feature set provided by the challenge organizers, we extract LPQ features that served as baseline in AVEC 2013 and are previously shown to be useful for facial action detection [209]. We detect and track the faces via a freely available facial landmark detector tool developed by Xiong and De La Torre [210]. Using the 49 landmark points provided by the detector we determine the facial area, align it using the eye centers, then crop and scale it such that the inter-ocular distance is 120 pixels and the right eye center is located at (65, 80) coordinates.

We extract LPQ features, focusing on only 5 regions: two containing eyes with brows and three covering the mouth area. We augment the raw LPQ descriptors with first and order second deltas.

5.2.2. Continuous Emotion Prediction Experiments on AVEC 2014 Corpus

In our experiments we adhere to the challenge protocol. We use the training set and optimize our model hyper parameters on the development set. It is important to note that all our experiments were carried out via training and predicting on the same task (e. g. train and predict on the Freeform task). No cross task learning is tested. When a viable method and hyper-parameter set is obtained, we re-train a model using the training and the development set to predict the independent test set, the labels of which are sequestered. The sub-challenge measure for ASC is Pearson's correlation.

In the ASC, the target variables are continuous and the challenge requires casting a prediction at 30Hz, which is the frame rate of video. Therefore, the video baseline features are found more suited for the task. The video modality baseline scores for the ASC are given in Table 5.9. Note that compared to AVEC 2013 [5], the baseline scores for the ASC are dramatically higher. This is partly due to employing multiple annotators and averaging the scores, and partly due to LGBP-TOP descriptors.

Table 5.9. AVEC 2014 Challenge video modality baselines (Pearson's Correlation Coefficient computed over all sequences).

Partition	Arousal	Valence	Dominance	Average
Devel.	0.412	0.355	0.319	0.362
Test	0.2062	0.1879	0.1959	0.1966

Since the ASC requires the prediction of continuous affective labels per video frame, to overcome the severe negative effects of undetected faces we carry out smoothing over the prediction contours after setting the training set mean of relevant affective dimension as rough prediction for the undetected frame. Mean smoothing is carried out on predictions of each clip over a window of 2K frames with respect to frame of interest.

As our motivation is to divide and conquer the data using ensemble CCA, we first show that ensemble averaging provides better results than using the whole set of features. This is intuitive as the error resulting from variance of predictors in known

to decrease via combining multiple learners [28]. The feature level fusion performance of four regions of interest on the development set is given in Table 5.10. The results are given in Pearson's Correlation Coefficient (PCC) that is the challenge measure for ASC. We observe that the smoothing has a major effect on performance, however the results only reach the development set baseline even when smoothed with K=120.

Table 5.10. Performance of CCA correlate of four inner regions of the baseline video features. Here, all the regions are combined at feature level.

PCC	Arousal	Valence	Dominance	**Avg**
Correlate	0.3085	0.3626	0.2855	**0.3189**
Smth. Correlate	0.3503	0.4187	0.3375	**0.3688**
1-NN Pred. On Correlate	0.2841	0.3282	0.2515	**0.2879**
Smth. 1-NN Pred. On Correlate	0.3525	0.4207	0.3356	**0.3696**

The development set performance of CCA ensembling can be seen in Table 5.11. We see that when projected separately, even a simple mean combination outperforms the baseline. Combined with smoothing, it is possible to reach an average PCC of 0.4066 using CCA regression and 0.4073 when 1-Nearest Neighbor regressor is used on the extracted CCA covariate. Since the performance difference of regressors in smoothed predictions is negligible and without smoothing the former performs better, we report further results with only CCA based regression.

Table 5.11. Performance of CCA correlates of four inner regions projected separately. K=120 is used for smoothing (Smt.).

PCC	Arousal	Valence	Dominance	**Avg**
Mean Correlate	0.4569	0.3916	0.2748	**0.3744**
Smth. Mean Correlate	0.4984	0.4201	0.3012	**0.4066**
1-NN Pred. On Mean Correlate	0.4012	0.3444	0.2385	**0.3280**
Smth. 1-NN Pred. On Mean Correlate	0.4987	0.4206	0.3025	**0.4073**

We next add the LPQ features into the loop and carry out tests on the development set. We use all the features from 5 aforementioned regions in LPQ, since individually they were not found to yield good performance. Although LPQ features

do not provide as good individual performance as LGBP-TOP, they contribute to overall performance in ensemble setting (see Table 5.12). We see both the non-smoothed and smoothed performance increase yielding a maximum PCC performance of 0.449.

Table 5.12. Using CCA regression ensemble of LPQ features (1 covariate) and LGBP-TOP features (4 regional covariates). K=150 is used for smooting.

	Arousal	Valence	Dominance	**Avg.**
No smoothing	0.4636	0.3992	0.3256	**0.3961**
Smoothing	0.5161	0.4424	0.3871	**0.449**

The computation of PCC over all sequences regardless of the task gives a result of 0.449. We see that the results change dramatically when the two tasks are handled separately. Using K=150 for smoothing on LPQ plus LGBP-TOP features, the PCC result in the Northwind task is 0.538, while in the Freeform the same setting gives a PCC performance of 0.423. The dramatic difference stems from the nature of the Freeform task: since this is a more in-the-wild task there is a large variation in terms of spoken content and type of reactions. Moreover, difference between the average of two task-dependent correlations (0.481) and the PCC computed over all sequences (0.449) is attributed to difference of variances in two tasks.

Figure 5.3 shows the effect of smoothing on PCC performance of the Freeform and Northwind tasks, respectively. Both tasks use CCA ensemble in the same setting given in Table 5.12. The figures also show that the effect of smoothing is mostly salient with smaller values of K, where the slope of contribution of smoothing decreases with increasing K. When we wish to build real-time emotion recognition systems, a small value of K (e. g. in range 10-30), would be appropriate considering the performance-efficiency dilemma.

5.2.2.1. Enhanced Visual System for Test Set. Motivated by the performance of feature partitioning based CCA ensembles with simple averaging as fusion rule, we extend our test set system in three ways. First, for computational issues (e. g. singularity and time complexity) we use Principal Component Analysis (PCA) prior to CCA. Second,

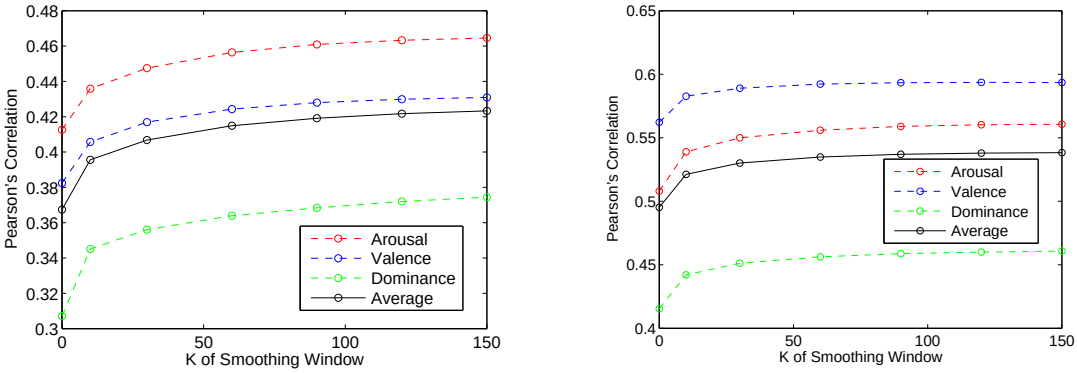

Figure 5.3. Development set performance with respect to K of smoothing on the Freeform (left) and Northwind (right) tasks

we extend feature level ensemble to instance level ensemble by learning CCA projection bases from the training and development sets then combining them in a fusion matrix (FM) similar to partition based Ensemble CCA [211]:

$$\mathbf{FM_t} = \begin{bmatrix} \mathbf{X^{tr}}\widehat{\mathbf{W}}_\mathbf{t}^{\mathbf{tr}} + \mu_\mathbf{t}^{\mathbf{tr}} & \mathbf{X^{tr}}\widehat{\mathbf{W}}_\mathbf{t}^{\mathbf{dev}} + \mu_\mathbf{t}^{\mathbf{dev}} \\ \mathbf{X^{dev}}\widehat{\mathbf{W}}_\mathbf{t}^{\mathbf{tr}} + \mu_\mathbf{t}^{\mathbf{tr}} & \mathbf{X^{dev}}\widehat{\mathbf{W}}_\mathbf{t}^{\mathbf{dev}} + \mu_\mathbf{t}^{\mathbf{dev}} \end{bmatrix}, \tag{5.1}$$

where t denotes the target affective dimension, superscripts "tr" and "dev" correspond to "training" and "development", respectively. For simplicity, the linear projection terms in Equation 2.11 are combined in Equation 5.1: $\widehat{\mathbf{W}} = \mathbf{W}\mathbf{V}^\dagger$. The third improvement over the development system is that the fusion matrix is stacked to a second level CCA to learn optimal projection weights, instead of simple averaging. The overall test system pipeline is given in Figure 5.4.

On the challenge test set we get PCC scores of 0.3915, 0.3837, and 0.4043 for arousal, valence and dominance, respectively; reaching and average of 0.3932, which outperforms the challenge test set baseline 100%, relative.

5.2.3. Overview

In this section, we utilize CCA to extract affective covariates in visual modality. We further employ CCA as a regressor and combine regional facial features in ensemble setting. The ensemble is obtained by training the most informative facial regions

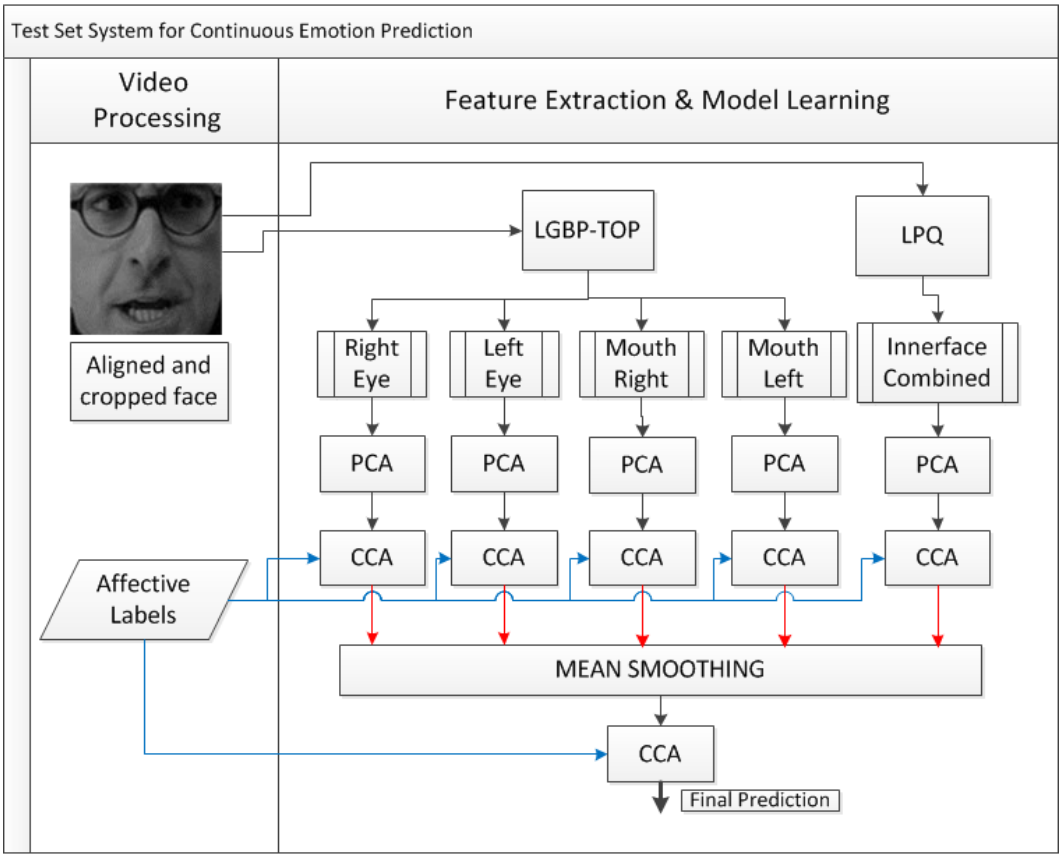

Figure 5.4. Illustration of continuous emotion prediction system. The predictions from regional Canonical Correlation Analyzers of the training and development sets are stacked to a second level CCA. For simplicity, 2-fold learning is not shown.

(corresponding to eyes and mouth area) separately. We show that CCA ensembling improves over mere feature fusion, and the predictions can be dramatically improved via mean smoothing. We also introduce AVEC 2013 baseline features to accompany the AVEC 2014 video baseline set. We observe that LPQ features are not individually sufficient but improve prediction performance collectively. For the ASC test set predictions, ensemble CCA is extended to instance space partitioning in addition to feature space partitioning, reaching an overall PCC score of 0.3932, which outperforms the challenge baseline PCC of 0.1966, dramatically.

5.3. Multimodal Prediction of Depression Severity Level

Depression can be defined as a state of low mood and aversion to activity that can affect a person's thoughts, feelings, behaviors, and sense of well-being. In this chap-

ter, we present proposed methods for audio and video based prediction of depression severity level using Audio-Visual Emotion Challenge (AVEC) 2013 and 2014 protocols [5, 16]. Both challenges use portions of the same depression corpus (AVDLC) that is introduced in Section 5.2.1. In AVDLC, the participating individuals are recorded via a web cam while they are guided by a presentation to do various tasks such as singing, reading and counting. The level of depression is measured by Beck-Depression Index II (BDI-II), a 21 item multiple-choice inventory [207]. The organizers provided baseline video and acoustic features along with the video clips partitioned into training, developments and test sets.

We propose different methods to handle features of each challenge. On AVEC 2013, we use CCA as feature selector in audio and as covariate extractor in video modality. On AVEC 2014, we use ELM to model both audio and video features.

5.3.1. Experiments with AVEC 2013 Corpus

The best test set RMSE results reported on AVEC 2013 Corpus/DSC are listed in Table 5.13. We see that the best reported score is obtained using audio only modality.

Table 5.13. Best test set results reported on AVEC 2013/DSC.

Work	Modality	RMSE
Cummins *et al.* [212]	Audio	10.17
Meng *et al.* [213]	Audio-visual	10.96
Cummins *et al.* [214]	Audio	11.37

5.3.1.1. Baseline Feature Sets. The acoustic feature set is the same as the one given in Section 5.2. The features are computed on short episodes of audio data. Since the Challenge dataset contains long continuous recordings, three segmentations have been performed: (i) voice activity detection (VAD) based, (ii) overlapping short fixed length segments (3 seconds) and, (iii) overlapping long fixed length segments (20 seconds).

For VAD segmentation, pauses of more than 200 ms are used to split speech activity segments. In short and long segmentation, the windows are shifted forward at a rate of one second. Functionals are then computed over each segment. Together with the per instance computation of functionals, the baseline feature set is provided in 4 versions to grasp relatively short-long acoustic characteristics of speech intended for depression and affect tasks. See Table 2 for the distribution of instances.

Table 5.14. Instance distribution per partition and segmentation.

#	Train	Dev	Test
Per Clip	50	50	50
VAD Seg	6015	5763	5946
Short Seg	23863	23513	23824
Long Seg	23439	23087	23399

The baseline video features consist of geometric features (e. g. head pose and coordinates) and Local Phase Quantization features (LPQ). Since the LPQ histogram representation does not keep the structural information of facial features, the face is divided into $4 \times 4 = 16$ regions and an LPQ histogram is computed per region. On the overall, the baseline LPQ feature set provides $16 \times 256 = 4096$ features for each face detected frame.

5.3.1.2. Experimental Results for Acoustic Depression Prediction. We tackle the acoustic depression severity level prediction problem via feature selection. We propose three CCA based feature selection methods introduced in Section 3.1, namely SLCCA-Filter, mRMR-CCA and MCR-CCA comparing their performance with CFS [121].

In our experiments, we used the WEKA [184] implementation of CFS with "Best First" search and Bagging-REPTree (BRep) from the same package as classifier. The hyper-parameters of both methods are left as default. As detailed before, we followed the training, development and testing protocol of the challenge. Therefore, we optimized the investigated feature selection methods on the development set and finally

used the optimal setting for predicting labels on the sequestered test set. For developing candidate hypotheses and selecting the best features for challenge test set, we utilized the training and development set. Baseline acoustic features using Support Vector Machine Regressor (SVR) with linear kernel gives Mean Absolute Error (MAE) of 8.66 and Root Mean Square Error (RMSE) of 10.75 for the development set [5].

We first used SVR (Linear Kernel, τ=0.0001) and Bagging REPTree (V = 0.001) as classifiers on VAD segmented features. In all our experiments we tested five feature settings: (i) All Baseline Features (denoted All) together with selected sets using (ii) SLCCA-Filter (iii) mRMR-CCA (iv) MCR-CCA and as independent benchmark (v) Correlation Based Feature Selection (CFS). Over five feature settings, we obtained better results with BRep (mean RMSE 11.15) against SVR (mean RMSE 12.24) with an order-of-magnitude less training time. So, we choose BRep as regressor.

We next experimented with five feature settings in all four segmentations. Considering the computational complexity of CCA (whose bottleneck is inversion of covariance matrix of samples, which scales cubicly with the number of selected features) we used the first 100 ranked features for MCR-CCA and mRMR-CCA. Once the threshold is determined, the number of features for SLCCA-Filter is automatically determined after a single application of CCA between the whole feature set versus the continuous depression labels. To probe the SLCCA-filter performance, we tested a set of thresholds $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$. The best development set results were obtained with a threshold of 10^{-5}.

A summary of experiments is given in Table 5.15. In accordance with the results reported in [5], we observe that the long segmentation provides the best results for the depression task. Moreover, the simplest CCA based selection method, namely SLCCA-Filter, yields the best RMSE results in all segmentations. The number of selected features with SLCCA-Filter ranges from 387 (long seg) to 467 (short seg) with segmented sets. However, due to increased nullity of covariance (with 50 samples as opposed to 2 268 dimensions) in per instance set, the number of features contributing to covariate is found to be 49.

Table 5.15. Development set performances per feature setting and segmentation.

	VAD SEG		**SHORT SEG**	
	MAE	**RMSE**	**MAE**	**RMSE**
All	9.01	11.25	8.37	10.42
SLCCA-Filter	9.13	11.15	7.99	10.36
MCR-CCA	9.31	11.54	8.73	10.86
mRMR-CCA	9.47	11.48	8.93	10.83
CFS	9.31	11.42	8.55	10.57
	LONG SEG		**PER CLIP**	
	MAE	**RMSE**	**MAE**	**RMSE**
All	7.93	10.24	9.75	11.89
SLCCA-Filter	7.84	10.22	8.92	11.00
MCR-CCA	8.27	10.72	9.81	11.55
mRMR-CCA	8.80	10.98	9.11	11.01
CFS	8.24	10.22	9.30	11.46

Focusing on long segmentation, we tested the first 400 ranked features for MCR-CCA and mRMR-CCA. The results were not found to improve considerably over the first 100 features: 8.13 MAE, 10.3 RMSE for MCR-CCA and 8.40 MAE and 10.97 RMSE for mRMR-CCA. Moreover, union and intersection of the selected feature sets pairwise did not improve over the performance of the SLCCA-Filter, individually.

We therefore used SLCCA-Filter method with long segmentation to train a model for the challenge test set. We obtained 7.83 MAE and 9.78 RMSE, improving challenge baseline test set RMSE performance (14.12) 30%, relative. These results also compare favorably to the best test set result of Meng *et al.* [213] (10.96 RMSE using audio-visual fusion) and Cummins *et al.* [212] (10.17 RMSE by using only audio information). Interestingly, unlike these recent studies that report better development set results using more complex systems, the development set performance of our computationally efficient SLCCA-Filter system is highly indicative of test set performance. Thus, SLCCA-Filter is thought to achieve the intended goal of avoiding over-fitting.

5.3.1.3. Experimental Results for Audio-Visual Depression Prediction. In this part, we use CCA for feature extraction in video and audio-visual modalities. We fuse the regional video features with the selected acoustic features that give the best performance in the previous subsection. To focus on the discriminative power of the extracted covariate, we use 1-Nearest Neighbor (1-NN) as regressor, and then average the predictions over the clip to get a final score. When individual regression performances of fused features are compared, the best results are consistently observed with regions corresponding to right eye, left eye and right half of the mouth. The stillness of the eyes, measured with the low variance of features from these regions, is found to be correlated with depression. Therefore, eyes indicate depressive mood.

We obtain canonical covariates from regional video features. For this, we remove the training set mean of each regional LPQ features (mode plus range) and apply the projection to development set. We then apply 1-NN regressor on the resulting 16-dimensional covariate space and obtain 6.90 MAE, 8.61 RMSE on the development set. If we combine the decisions of regional regressors instead, we get 7.61 MAE, 9.16 RMSE. When only the best 6 regional regressors are combined, the performance barely reaches the feature level combination: 7.07 MAE, 8.56 RMSE.

We fuse the selected 387 acoustic features with each of regional visual features using CCA. Since CCA provides projections from either views, i. e. video and audio features, we utilize the video projections. The number of covariates in CCA is limited to the minimum of matrix ranks of two views. Since the selected audio features are smaller in number (387 as opposed to 512 video features) and they are linearly independent, the maximum number of covariates is 387. We apply a second level CCA between $p = \{50, 100, 150, 200, 250\}$ covariates from each region and the target labels to obtain a final single covariate for regression. The RMSE performance of regional regressors with varying number of covariates are given in Figure 5.5. We observe that the best three regional regressors are always number 6, 7 and 10, which correspond to left eye, right eye and the right part of the mouth, respectively. This is intuitive as the most potent parts of the face for action recognition are eyes and mouth area. This preliminary result helps reduce the computations to one quarter by focusing on the

inner 2-by-2 square in the 4-by-4 partitioning of facial image. Moreover, the RMSE is observed to decrease up to 200 covariates but does not improve further.

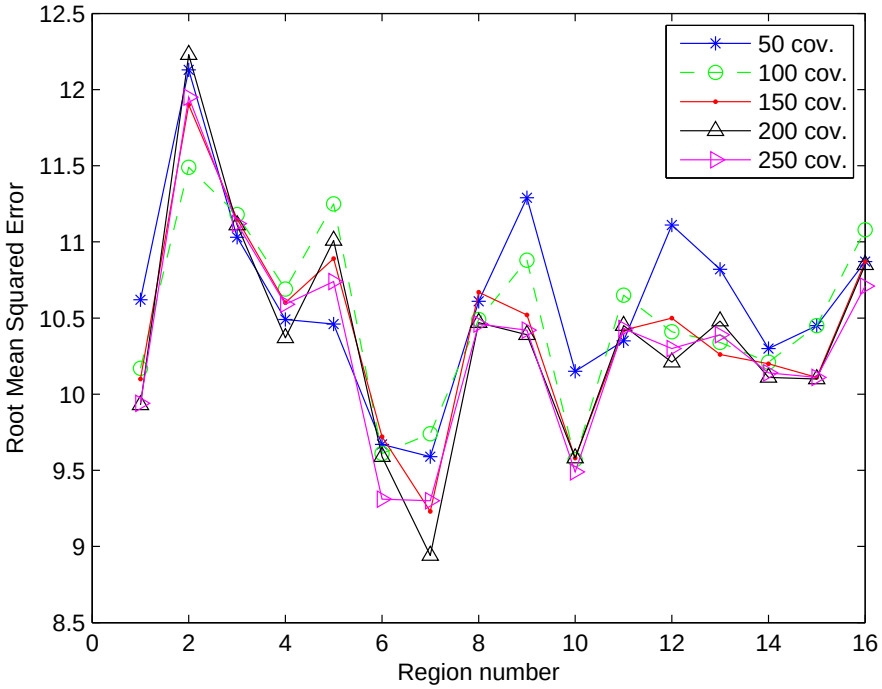

Figure 5.5. Canonical correlations of regional video features vs. depression labels in the training set.

We finally concatenate all the local audio-visual covariates into a feature vector and apply CCA against the target labels to obtain a single depression covariate. The best RMSE results were obtained when 100 regional audio-visual covariates are combined (i.e. 1600 features) for second stage CCA in a similar fashion to Ensemble CCA [211]. An interesting finding is that the development set correlation of audio-visual covariates is higher than the training set. With 100 local covariates, the training set ρ is found as 0.930 and development set ρ is 0.952. For video modality ρ reduces from 0.948 to 0.636, whereas in audio modality the decrease is more dramatic (from 0.794 to 0.306).

For the test set we retrain training and development sets for the video-only model with 16 regional covariates (M1), video-only model using only 3 best performing regions (M2), audio-visual fusion with 150 covariates (M3), and decision fusion of the best

acoustic model (M0) with M2 (M4). The results given in Table 5.16 indicate that audio-visual feature level fusion does not improve over video only features, while audio-visual decision fusion provides the best result. Note that all results are better than the challenge baselines.

Table 5.16. Challenge test set results on AVEC 2013 DSC.

Model	Modality	MAE	RMSE
M0	Audio	7.83	9.78
M1	Video	7.97	9.94
M2	Video	7.86	9.72
M3	Audio-visual (feature level)	8.79	10.81
M4	Audio-visual (decision fusion)	7.68	9.44

Considering scores given in Table 5.13, our best test set score using 1-NN regressor achieves the state-of-the-art. The regressor and its hyper parameter were not optimized in this work, in order to sharpen the effect of proposed CCA based feature reduction approaches. The study can be improved by employing more potent visual descriptors and more sophisticated regressors.

5.3.1.4. Overview. In this section, we employed three novel CCA based feature selection methods to reduce the massive dimensionality observed with the state-of-the-art acoustic feature sets in affective computing. Results revealed that the computationally simple CCA based feature selection method worked the best on the development set. Using only 17% of original features, the SLCCA-Filter system yielded 30% decrease of RMSE over the baseline on challenge test set. We also used CCA to extract depression covariates in visual and audio-visual modalities. We see that the facial regions corresponding to eyes and mouth area provide the best regression scores. We observe that multimodal fusion yields more robust covariates in the development set, which may be important for affect recognition. The challenge test set scores indicate the efficacy of our simple approach. The best test set results that achieve the state-of-the-art on this corpus are obtained with audio-visual decision fusion.

5.3.2. Experiments with AVEC 2014 Corpus

As mentioned earlier, AVEC 2014 uses the same base corpus with AVEC 2013, with a focus on two tasks (called Freeform and Northwind) instead of 12. Note that this corpus is also used for continuous emotion recognition in Section 5.2. Therefore, details of the 2014 corpus and features can be found in Section 5.2.1.

In this section, we utilize Moore-Penrose Generalized Inverse (MPGI) and ELM for audio-visual depression recognition. As detailed in Section 2.2.1.4, ELMs provide a unified framework for regression and multi-class classification for a generalized Single Layer Feedforward Networks (SLFNs) including Support Vector Machines (SVMs) [14]. We use Principal Component Analysis (PCA) for the input layer instead of random projections used in basic ELM [13]. Unlike ELMs, we do not utilize a non-linear activation function at hidden layer. For the output layer, the learning rule is the same with basic ELM that provides a least squares solution.

5.3.2.1. Experimental Results for Audio-Visual Depression Prediction. For the DSC video sub-system, we carry out a slight modification on the ASC pipeline given in Section 5.2.2.1. We keep frame level learning as is taking into account only features from face detected frames. We replace first level CCA by MPGI and second level CCA by simple averaging. In our preliminary experiments, we observe that MPGI yields better RMSE performance on the development set than CCA. To our surprise, simple averaging is also found to give better RMSE performance compared to CCA based fusion for final stage. Different from ASC, the predictions from both the Northwind and Freeform tasks can be combined to give a final score for the clip. We therefore train PCA + MPGI based ensemble ELM systems per task, then evaluate their individual and combined (averaged) performance.

For audio sub-system, we utilize long segmented baseline acoustic feature sets, since this segmentation is shown to suit depression sub-challenge [5, 9]. We used ELMs with Linear Kernel and optimized the complexity parameter in the range $2^{\{-15,-14,\ldots,0\}}$.

Here also, we trained separate models on the training and development sets in a 2-Fold cross validation manner. Prior to model learning acoustic features are normalized such that they are in the range [-1,1]. We observe that models learned from the Freeform task does not yield better performance than challenge baseline. So, we used only predictions from the Northwind task in audio sub-system. Averaging models learned from the training and development sets, we obtain our acoustic system's final output. For audio-visual fusion, we combine the Northwind audio and video sub-systems for consistency of the video task. The sequestered test set scores of five systems are listed in Table 5.17. The results indicate that audio only modality generalizes better than video only modality, that was found to give lower than 10 RMSE on the development set. The best results are obtained with audio-visual decision fusion.

Table 5.17. AVEC 2014 Challenge test set scores of five systems for DSC.

Task	Modality	MAE	RMSE
Freeform	Video	8.284	10.519
Northwind	Video	8.254	10.315
Northwind+Freeform	Video	8.202	10.269
Northwind	Audio	7.962	9.978
Northwind	Audio-visual	7.693	9.611

5.3.2.2. Overview. In this section, we utilize audio-visual fusion of ELM based systems for predicting depression severity level. The visual ensemble system proposed for continuous emotion prediction in Section 5.2.2.1 is also used in audio-visual depression recognition, where CCA is replaced with Moore-Penrose generalized inverse to find a least squares solution with minimum L_2 norm projection. For audio based depression prediction, we employ ELMs with Linear kernel. Combining audio and video subsystems learned from the Northwind task, we reach a test set RMSE score of 9.611 improving the challenge baseline RMSE score of 10.859. In the part, we did not use any feature selection method, which we think might be useful especially for audio systems. In our future studies, we are planning to apply variants of a recently introduced multi-view feature selection approach [10] utilizing domain knowledge to partition the feature set.

6. CONCLUSIONS

Throughout the thesis study, which has grown around speech emotion recognition, a set of paralinguistic and affective computing tasks are dealt with. The reason of starting from emotions is that all paralinguistic tasks are inherently related to them [2] and have the same processing pipeline. The bottlenecks of the pipeline comprised the research sub-problems of this thesis. In the next section, thesis contributions that target these bottlenecks are summarized. Section 6.2 concludes with overall discussion of the thesis and future directions.

6.1. Summary of Thesis Contributions

The contributions of the thesis can be summarized as follows:

(i) Novel discriminative projection based feature selection methods: A set of novel feature selection methods based on Canonical Correlation Analysis (CCA) are proposed to deal with curse of dimensionality in the paralinguistics processing pipeline. These are Samples versus Labels CCA Filter (SLCCA-Filter), its randomized version (SLCCA-RAND), its multi-view version that uses domain knowledge (SLCCA-LLD), Minimum Redundancy Maximum Relevance (mRMR) CCA and Maximum collective Relevance (MCR) CCA. These filters are applied on recent paralinguistic challenge corpora, resulting in state-of-the-art performance in depression severity level prediction (SLCCA-Filter), acoustic physical load (heart pulse) recognition (SLCCA-LLD) and acoustic conflict recognition (SLCCA-RAND). These filters are also used to improve system performance in audio-visual systems for robust affective computing.

(ii) A novel automatic model selection algorithm for Mixtures of Factor Analyzers: A new model selection algorithm is proposed for statistical unsupervised modeling of large datasets via mixtures of factor analyzers. The proposed approach is much faster than Monte Carlo based fully Bayesian alternatives (minutes to hours of training instead of weeks on an ordinary PC), and much more accurate and

parsimonious than the traditional Gaussian mixture model selection techniques in the literature. The superiority of the approach is shown on multiple machine learning datasets, and the method is applied to child emotion recognition in natural conditions.

(iii) Cascaded Normalization for paralinguistic feature preprocessing: The cascaded normalization approach was used for the first time for combining speaker level normalization with non-linear normalization and instance level normalization. The proposed approach boosted the performance on the eating condition sub-challenge of INTERSPEECH 2015 paralinguistic challenge.

(iv) Adaptation of recent machine learning paradigms to paralinguistic problems: A set of classifiers are employed for the first time in several paralinguistics tasks, giving state-of-the-art or competitive results. The thesis incorporates the first application of random forests to laughter detection, the first application of extreme learning machines to audio-visual emotion recognition in the wild and depression severity prediction, and the first application of canonical correlation analysis to continuous emotion recognition from video. Similarly, a recently popular video modeling method, the Fisher Vector encoding, is employed in audio domain for acoustic feature modeling over the speech utterance for the first time, giving a marked improvement over the state-of-the-art acoustic systems.

6.2. Discussion and Future Directions

It is important to note that all experiments on paralinguistic and affective computing are carried out using the recent challenge corpora: INTERSPEECH ComParE 2013-2015, AVEC 2013-2014, EmotiW 2014, and MAPTRAITS 2014. We adhere to the standard challenge protocols, where the test labels are sequestered. Participating in challenges and/or using challenge corpora gave us several opportunities including i) repeatability & comparability of studies ii) measuring generalization capability of models on unseen data iii) engineering on novel and challenging problems iv) competing with and learning from other teams leading the field.

This PhD thesis primarily aimed to identify and resolve bottlenecks of the state-

of-the-art processing pipeline by proposing/employing efficient and accurate machine learning methods [215]. We targeted the adverse effects of brute-forced feature extraction, which is popularly used in all paralinguistic systems and challenges. To overcome the curse of dimensionality, we proposed several discriminative projection based feature selection methods and strategies, whose efficacy are validated with the aforementioned challenge corpora for acoustic depression severity level prediction [9], acoustic conflict recognition [11], and acoustic physical load prediction [10].

The advantage of using CCA in feature selection is twofold. First, in supervised setting it can be applied in both regression and classification, making it superior to LDA variants. As long as CCA based supervised feature selection is concerned, we observed that using regression labels provide better feature ranking compared to their discretized (categorical) labels [11]. This is attributed to better mapping of continuous feature space to continuity in the target space and loss of information during discretization. Second, CCA can be used in an unsupervised (as in the multi-view case) and/or semi-supervised setting [100, 118, 119]. In this thesis, we used CCA for supervised feature selection, while semi-supervised feature reduction remained as future work.

We employed CCA for a set of purposes, in addition to its common use for covariate extraction. We showed that discriminative projection matrices can be used for feature ranking, whose output generalizes better to unseen data than the projection itself, especially for acoustic features. In video modality, where individual features are insufficient for prediction and are similar in their (first and second order) statistics (unlike acoustic suprasegmental features), CCA is employed as covariate extractor [24] and regressor [19].

For speech emotion recognition, it is known that the acoustic content favors arousal classification, while valence is hard to classify due to weak cues. For better discrimination of valence related affect, other modalities (such as linguistics or video) are employed in the literature. In our studies, we used audio-visual decision fusion and obtained improved performance for emotion recognition in-the-wild [17, 18] and for predicting depression severity level [19, 24].

In two audio-visual corpora, we observed higher performance with features from the inner facial regions. While in AVEC 2013/2014 these regions correspond to the eyes and mouth area, in EmotiW 2014 the inner regions do not cover eyes and mouth fully. This implies that the higher performance of inner facial regions is mostly due to less sensitivity to occlusions and registration errors.

Apart from multimodal fusion, we obtained improved performance with smoothing for tasks involving continuous prediction in space and time [19,25,26]. An important note here is, when the predictions are incorrect (e. g. negatively correlated with the ground truth), smoothing does not have a corrective but intensifying effect [26].

Working on affective data, an important problem is the subjectivity of annotations. To avoid the label uncertainty, it is common to employ multiple annotators. 2014 versions of the AVEC and EmotiW emotion challenges provided very similar data with increased number of annotators compared to their 2013 versions. Increasing the number of annotators not only elevated the baseline scores, but also boosted the top performances reached in the respective challenges, highlighting the importance of further research on annotator modeling.

To cope with large acoustic variability due to speakers, recording conditions and the spoken content, Gaussian Mixture Models are widely employed in speech processing. However, model selection issue is generally omitted. We suggest the use of Mixture of Factor Analyzers, which is capable of modeling a much wider range of models between diagonal and full covariance Gaussian with compact parametrization. We propose an efficient and deterministic MoFA (AMoFA) algorithm that simultaneously handles model selection and parameter estimation. The proposed AMoFA algorithm uses incremental-decremental model adaptation to best fit the complexity of the model to data complexity. We applied AMoFA to model LLDs of affective child speech in natural conditions and obtained better results compared to SVM modeling of suprasegmental openSMILE feature set that is used in the latest paralinguistic challenges. Further studies for application of AMoFA to affective computing will be focused on recent challenge corpora, on which we have previously applied discriminative modeling.

With an aim of finding a more robust feature representation compared to the state-of-the-art suprasegmental openSMILE features, we employed the Fisher Vector (FV) encoding for utterance modeling [22]. We found that speaker normalization on the openSMILE features provides a marked improvement for recognition of the speaker state. Therefore, we first used the FV encoding to alleviate speaker variability, then implemented speaker normalization after speaker clustering. Using only the FV encoding, it was possible to reach a performance improvement equal to that obtained by speaker normalization on the openSMILE features. Applying a cascaded normalization scheme on the FV, we outperformed the challenge test set baseline by a large margin. As a future work, the methodology employed therein will be applied to other paralinguistic and affective computing tasks, such as cross-corpus acoustic emotion recognition.

This thesis study did not employ deep learning, which is increasingly popular in signal processing and machine learning. Instead, we employed ELMs, which learn extremely fast and provide accurate predictions. Our experiments on audio-visual emotion recognition in-the-wild [17] showed that the proposed ELM based system that use only 4 modality-specific sub-systems reaches the performance of the top system that is composed of 25 sub-systems, which extensively use deep learning [114].

Future works for combining and extending the novel methods proposed in this thesis are as follows. First, it is possible to implement multi-cluster extension CCA based feature filters using AMoFA, by exploiting the relationship between MoFA and Mixtures of Probabilistic CCA. Next, CCA can be used to initialize AMoFA for multimodal, multi-view data. We have illustrated efficacy of AMoFA in clustering and classification tasks. It is also possible to employ AMoFA as a locally linear, globally non-linear feature extractor. As a future work, we plan to design a fast learning neural network (NN) with automatic system identification (determination of hidden node number) capability. The first layer projection of this NN can be learned by AMoFA, and the second layer projection can be efficiently learned via Moore-Penrose Generalized Inverse (MPGI), which is used in ELMs. This approach can be successfully applied to regression problems, where classical NNs fail in the presence of multiple clusters, since they expect continuous functions to approximate.

REFERENCES

1. Schuller, B., *In Salah, A. A. and Gevers, T. (eds) Computer Analysis of Human Behavior*, chap. Voice and Speech Analysis in Search of States and Traits, pp. 227–253, Springer, 2011.

2. Cowie, R., N. Sussman and A. Ben-Zeev, *Emotion-Oriented Systems: The Humaine Handbook*, chap. Emotion: Concepts and Definitions, pp. 9–32, Springer, 2011.

3. Schuller, B., B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth and G. Rigoll, "Cross-corpus Acoustic Emotion Recognition: Variances and Strategies", *IEEE Transactions on Affective Computing*, Vol. 1, No. 2, pp. 119–131, 2010.

4. Schuller, B., S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism", *Proceedings of INTERSPEECH*, pp. 148–152, ISCA, Lyon, France, 2013.

5. Valstar, M., B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie and M. Pantic, "AVEC 2013–The Continuous Audio/Visual Emotion and Depression Recognition Challenge", *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, pp. 3–10, 2013.

6. Schuller, B., S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load", *Proceedings of INTERSPEECH*, pp. 427–431, Singapore, 2014.

7. Bishop, C. M., *Pattern Recognition and Machine Learning*, Vol. 4 of *Information*

Science and Statistics, Springer New York, 2006.

8. Kaya, H. and A. A. Salah, "Adaptive Mixture of Factor Analyzers", *(submitted)*, 2015.

9. Kaya, H., F. Eyben, A. A. Salah and B. W. Schuller, "CCA Based Feature Selection with Application to Continuous Depression Recognition from Acoustic Speech Features", *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, pp. 3757–3761, Florence, Italy, 2014.

10. Kaya, H., T. Özkaptan, A. A. Salah and S. F. Gürgen, "Canonical Correlation Analysis and Local Fisher Discriminant Analysis based Multi-View Acoustic Feature Reduction for Physical Load Prediction", *Proceedings of INTERSPEECH*, pp. 442–446, ISCA, Singapore, 2014.

11. Kaya, H., T. Özkaptan, A. A. Salah and S. F. Gürgen, "Random Discriminative Projection based Feature Selection with Application to Conflict Recognition", *IEEE Signal Processing Letters*, Vol. 22, No. 6, pp. 671–675, 2015.

12. Wold, H., "Partial Least Squares", S. Kotz and N. L. Johnson (Editors), *Encyclopedia of Statistical Sciences*, pp. 581–591, Wiley New York, 1985.

13. Huang, G.-B., Q.-Y. Zhu and C.-K. Siew, "Extreme Learning Machine: Theory and Applications", *Neurocomputing*, Vol. 70, No. 1, pp. 489–501, 2006.

14. Huang, G.-B., H. Zhou, X. Ding and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 42, No. 2, pp. 513–529, 2012.

15. Dhall, A., R. Goecke, J. Joshi, K. Sikka and T. Gedeon, "Emotion Recognition in the Wild Challenge 2014: Baseline, Data and Protocol", *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*, pp. 461–466,

2014.

16. Valstar, M., B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie and M. Pantic, "AVEC 2014–3D Dimensional Affect and Depression Recognition Challenge", *Proceedings of the 4rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, 2014.

17. Kaya, H. and A. A. Salah, "Combining Modality-Specific Extreme Learning Machines for Emotion Recognition in the Wild", *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pp. 487–493, 2014.

18. Kaya, H. and A. A. Salah, "Combining Modality-Specific Extreme Learning Machines for Emotion Recognition in the Wild", *Journal on Multimodal User Interfaces*, 2015, `http://dx.doi.org/10.1007/s12193-015-0175-6`, [Accessed May 2015].

19. Kaya, H., F. Çilli and A. A. Salah, "Ensemble CCA for Continuous Emotion Prediction", *Proceedings of the 4th ACM International Workshop on Audio/Visual Emotion Challenge*, pp. 19–26, ACM, Orlando, Florida, USA, 2014.

20. Perronnin, F. and C. Dance, "Fisher Kernels on Visual Vocabularies for Image Categorization", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Minneapolis, Minnesota, USA, 2007.

21. Perronnin, F., Y. Liu, J. Sánchez and H. Poirier, "Large-scale Image Retrieval with Compressed Fisher Vectors", *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3384–3391, 2010.

22. Kaya, H., A. A. Karpov and A. A. Salah, "Fisher Vectors with Cascaded Normalization for Paraliguistic Analysis", *(submitted)*, 2015.

23. Kaya, H., A. A. Salah, S. F. Gurgen and H. Ekenel, "Protocol and Baseline for Experiments on Bogazici University Turkish Emotional Speech Corpus", *Proceedings*

of 22nd IEEE Signal Processing and Communications Applications Conference (SIU 2014), pp. 1698–1701, 2014.

24. Kaya, H. and A. A. Salah, "Eyes Whisper Depression: A CCA based Multimodal Approach", *Proceedings of the 22nd International Conference on Multimedia*, ACM MM '14, pp. 961–964, Orlando, Florida, USA, 2014.

25. Kaya, H., A. M. Erçetin, A. A. Salah and F. Gürgen, "Random Forests for Laughter Detection", *Proceedings of Workshop on Affective Social Speech Signals*, 2013.

26. Kaya, H. and A. A. Salah, "Continuous Mapping of Personality Traits: A Novel Challenge and Failure Conditions", *Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge and Workshop*, MAPTRAITS '14, pp. 17–24, 2014.

27. Schuller, B., G. Rigoll and M. Lang, "Speech Emotion Recognition Combining Acoustic Features And Linguistic Information In A Hybrid Support Vector Machine - Belief Network Architecture", *ICASSP*, pp. 577–580, 2004.

28. Alpaydin, E., *Introduction to Machine Learning*, The MIT Press, Cambridge, 2nd edn., 2010.

29. Hermansky, H., "Perceptual Linear Predictive (PLP) Analysis of Speech", *The Journal of the Acoustical Society of America*, Vol. 87, No. 4, pp. 1738–1752, 1990.

30. Hermansky, H. and N. Morgan, "RASTA Processing of Speech", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 578–589, 1994.

31. Itakura, F., "Line Spectrum Representation of Linear Predictor Coefficients of Speech Signals", *Journal of Acoustic Society of America*, Vol. 57, No. S1, p. S35, 1975.

32. Bozkurt, E., E. Erzin, C. E. Erdem and A. T. Erdem, "Use of Line Spectral Frequencies for Emotion Recognition from Speech", *Proceedings of 20th International*

Conference on Pattern Recognition, pp. 3708–3711, 2010.

33. Bozkurt, E., E. Erzin, C. E. Erdem and A. T. Erdem, "Formant Position based Weighted Spectral Features for Emotion Recognition", *Speech Communication*, Vol. 53, No. 9-10, pp. 1186–1197, 2011.

34. El Ayadi, M., M. S. Kamel and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", *Pattern Recognition*, Vol. 44, No. 3, pp. 572–587, 2011.

35. Rabiner, L., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257–286, 1989.

36. Schuller, B., G. Rigoll and M. Lang, "Hidden Markov Model-based Speech Emotion Recognition", *Proceedings of IEEE International Conference on Multimedia and Expo*, Vol. 1, pp. 401–404, Los Alamitos, CA, USA, 2003.

37. Eyben, F., M. Wöllmer and B. Schuller, "OpenSMILE: The Munich Versatile and Fast Open-source Audio Feature Extractor", *Proceedings of the International Conference on Multimedia*, pp. 1459–1462, ACM, 2010.

38. Schuller, B., M. Lang and G. Rigoll, "Robust Acoustic Speech Emotion Recognition by Ensembles of Classifiers", *Fortschritte Der Akustik*, Vol. 31, No. 1, pp. 329–330, 2005.

39. Schuller, B., R. J. Villar, G. Rigoll and M. K. Lang, "Meta-Classifiers in Acoustic and Linguistic Feature Fusion-Based Affect Recognition", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)*, pp. 325–328, 2005.

40. Kächele, M., M. Glodek, D. Zharkov, S. Meudt and F. Schwenker, "Fusion of Audio-Visual Features using Hierarchical Classifier Systems for the Recognition of Affective States and the State of Depression", *Proceedings of the International*

Conference on Pattern Recognition Applications and Methods, pp. 671–678, 2014.

41. Ververidis, D. and C. Kotropoulos, "A State of the Art Review on Emotional Speech Databases", *Proceedings of 1st Richmedia Conference*, pp. 109–119, 2003.

42. Dhall, A., R. Goecke, S. Lucey and T. Gedeon, "Collecting Large, Richly Annotated Facial-Expression Databases from Movies", *IEEE MultiMedia*, Vol. 19, No. 3, pp. 34–41, 2012.

43. Amir, N., S. Ron and N. Laor, "Analysis of an Emotional Speech Corpus in Hebrew Based on Objective Criteria", *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pp. 29–33, 2000.

44. Schuller, B., R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application", *Image and Vision Computing*, Vol. 27, No. 12, pp. 1760–1774, 2009.

45. Slaney, M. and G. McRoberts, "BabyEars: A Recognition System for Affective Vocalizations", *Speech Communication*, Vol. 39, No. 3, pp. 367–384, 2003.

46. Fu, L., X. Mao and L. Chen, "Speaker Independent Emotion Recognition based on SVM/HMMs Fusion System", *IEEE International Conference on Audio, Language and Image Processing*, pp. 61–65, 2008.

47. Zhou, J., G. Wang, Y. Yang and P. Chen, "Speech Emotion Recognition Based on Rough Set and SVM", *The 5th IEEE International Conference on Cognitive Informatics. ICCI 2006.*, Vol. 1, pp. 53–61, IEEE, 2006.

48. Engberg, I. and A. Hansen, *Documentation of the Danish Emotional Speech Database (DES)*, Internal AAU Report, Center for Person Kommunikation, Denmark, 1996.

49. Burkhardt, F., A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, "A Database of German Emotional Speech", *Proceedings of INTERSPEECH*, pp. 1517–1520, 2005.

50. Martin, O., I. Kotsia, B. Macq and I. Pitas, "The eNTERFACE '05 Audio-Visual Emotion Database", *Proceedings of IEEE Workshop on Multimedia Database Management*, 2006.

51. Batliner, A., S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir and L. Kessous, "Combining Efforts for Improving Automatic Classification of Emotional User States", pp. 240–245, 2006.

52. Schuller, B., "Towards Intuitive Speech Interaction by the Integration of Emotional Aspects", *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, Vol. 6, p. 6, 2002.

53. Bänziger, T., M. Mortillaro and K. R. Scherer, "Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception", *Emotion*, Vol. 12, No. 5, pp. 1161–1179, 2012.

54. Kim, E. H., K. H. Hyun, S. H. Kim and Y. K. Kwak, "Speech Emotion Recognition Using Eigen-FFT in Clean and Noisy Environments", *Proceedings of the 16th IEEE International Symposium on Robot and Human interactive Communication (RO-MAN 2007)*, pp. 689–694, 2007.

55. Breazeal, C. and L. Aryananda, "Recognition of Affective Communicative Intent in Robot-Directed Speech", *Autonomous Robots*, Vol. 12, No. 1, pp. 83–104, 2002.

56. Liberman, M., K. Davis, M. Grossman, N. Martey and J. Bell, *Emotional Prosody Speech and Transcripts*, LDC2002S28, Philadelphia: Linguistic Data Consortium, 2002, `https://catalog.ldc.upenn.edu/LDC2002S28`, [Accessed May 2015].

57. Schuller, B., S. Reiter, R. Muller, M. Al-Hames, M. Lang and G. Rigoll, "Speaker

Independent Speech Emotion Recognition by Ensemble Classification", *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2005)*, pp. 864–867, 2005.

58. Morrison, D., R. Wang and L. C. De Silva, "Ensemble Methods for Spoken Emotion Recognition in Call-Centres", *Speech Communication*, Vol. 49, No. 2, pp. 98–112, 2007.

59. Douglas-Cowie, E., R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian and A. Batliner, "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data", *Proceedings of Affective Computing and Intelligent Interaction*, pp. 488–500, 2007.

60. Steininger, S., F. Schiel, O. Dioubina and S. Raubold, "Development of User-State Conventions for the Multimodal Morpus in SmartKom", *LREC Workshop on Multimodal Resources*, pp. 33–37, 2002.

61. Hansen, J. H., S. E. Bou-Ghazale, R. Sarikaya and B. Pellom, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database", *Proceedings of Eurospeech*, Vol. 97, pp. 1743–46, 1997.

62. Oflazoglu, C. and S. Yildirim, "Turkish Emotional Speech Database", *Proceedings of the 19th IEEE Conference on Signal Processing and Communications Applications (SIU)*, pp. 1153–1156, 2011.

63. Grimm, M., K. Kroschel and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database", *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 865–868, 2008.

64. Schuller, B., S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge", *Proceedings of INTERSPEECH*,

pp. 254–257, ISCA, Portland, OR, USA, 2012.

65. Schuller, B., S. Steidl and A. Batliner, "The Interspeech 2009 Emotion Challenge", *Proceedings of INTERSPEECH*, pp. 312–315, Brighton, UK, 2009.

66. Norman, W. T., "Toward an Adequate Taxonomy of Personality Attributes: Replicated Factor Structure in Peer Nomination Personality Ratings.", *The Journal of Abnormal and Social Psychology*, Vol. 66, No. 6, pp. 574–583, 1963.

67. Schuller, B., S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi and B. Weiss, "A Survey on Perceived Speaker Traits: Personality, Likability, Pathology, and the First Challenge", *Computer Speech & Language*, Vol. 29, No. 1, pp. 100–131, 2015.

68. Ivanov, A. and X. Chen, "Modulation Spectrum Analysis for Speaker Personality Trait Recognition", *Proceedings of INTERSPEECH*, pp. 278–281, 2012.

69. Montacié, C. and M.-J. Caraty, "Pitch and Intonation Contribution to Speakers' Traits Classification", *Proceedings of INTERSPEECH*, pp. 526–529, 2012.

70. Anumanchipalli, G. K., H. Meinedo, M. Bugalho, I. Trancoso, L. C. Oliveira and A. W. Black, "Text-Dependent Pathological Voice Detection", *Proceedings of INTERSPEECH*, pp. 530–533, 2012.

71. Brueckner, R. and B. Schuller, "Likability Classification-A Not so Deep Neural Network Approach", *Proceedings of INTERSPEECH*, pp. 290–293, 2012.

72. Buisman, H. and E. Postma, "The log-Gabor Method: Speech Classification using Spectrogram Image Analysis", *Proceedings of INTERSPEECH*, pp. 518–521, 2012.

73. Kim, J., N. Kumar, A. Tsiartas, M. Li and S. S. Narayanan, "Intelligibility Classification of Pathological Speech using Fusion of Multiple Subsystems", pp. 534–

537, 2012.

74. Lu, D. and F. Sha, "Predicting Likability of Speakers with Gaussian Processes", *Proceedings of INTERSPEECH*, pp. 286–289, 2012.

75. Huang, D.-Y., Y. Zhu, D. Wu and R. Yu, "Detecting Intelligibility by Linear Dimensionality Reduction and Normalized Voice Quality Hierarchical Features", *Proceedings of INTERSPEECH*, pp. 546–549, 2012.

76. Asgari, M., A. Bayestehtashk and I. Shafran, "Robust and Accurate Features for Detecting and Diagnosing Autism Spectrum Disorders", *Proceedings of INTERSPEECH*, pp. 191–194, 2013.

77. Gonzalez, D. M., D. Ribas, E. Lleida, A. Ortega and A. Miguel, "Suprasegmental Information Modelling for Autism Disorder Spectrum and Specific Language Impairment Classification", *Proceedings of INTERSPEECH*, pp. 195–199, 2013.

78. Lee, H.-Y., T.-Y. Hu, H. Jing, Y.-F. Chang, Y. Tsao, Y.-C. Kao and T.-L. Pao, "Ensemble of Machine Learning and Acoustic Segment Model Techniques for Speech Emotion and Autism Spectrum Disorders Recognition", *Proceedings of INTERSPEECH*, pp. 215–219, 2013.

79. Rasanen, O. and J. Pohjalainen, "Random Subset Feature Selection in Automatic Recognition of Developmental Disorders, Affective States, and Level of Conflict from Speech", *Proceedings of INTERSPEECH*, pp. 210–214, 2013.

80. Grèzes, F., J. Richards and A. Rosenberg, "Let Me Finish: Automatic Conflict Detection using Speaker Overlap", *Proceedings of INTERSPEECH*, pp. 200–204, 2013.

81. Gosztolya, G., R. Busa-Fekete and L. Tóth, "Detecting Autism, Emotions and Social Signals Using Adaboost", *Proceedings of INTERSPEECH*, pp. 220–224, 2013.

82. Schapire, R. E. and Y. Singer, "Improved Boosting Algorithms Using Confidence-rated Predictions", *Machine Learning*, Vol. 37, No. 3, pp. 297–336, 1999.

83. Busa-Fekete, R. and B. Kégl, "Fast Boosting Using Adversarial Bandits", *Proceedings of ICML*, pp. 143–150, 2010.

84. Sethu, V., J. Epps, E. Ambikairajah and H. Li, "GMM based Speaker Variability Compensated System for Interspeech 2013 Compare Emotion Challenge", *Proceedings of INTERSPEECH*, pp. 205–209, 2013.

85. Gupta, R., K. Audhkhasi, S. Lee and S. Narayanan, "Paralinguistic Event Detection from Speech Using Probabilistic Time-Series Smoothing and Masking", *Proceedings of INTERSPEECH*, pp. 173–177, 2013.

86. Janicki, A., "Non-linguistic Vocalisation Recognition Based on Hybrid GMM-SVM Approach", *Proceedings of INTERSPEECH*, pp. 153–157, 2013.

87. Van Segbroeck, M., R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos and S. S. Narayanan, "Classification of Cognitive Load from Speech using an i-Vector Framework", *Proceedings of INTERSPEECH*, pp. 751–755, Singapore, 2014.

88. Gosztolya, G., T. Grósz, R. Busa-Fekete and L. Tóth, "Detecting the Intensity of Cognitive and Physical Load Using Adaboost and Deep Rectifier Neural Networks", *Proceedings of INTERSPEECH*, pp. 452–456, 2014.

89. Kua, J. M. K., V. Sethu, P. Le and E. Ambikairajah, "The UNSW Submission to Interspeech 2014 ComParE Cognitive Load Challenge", *Proceedings of INTERSPEECH*, pp. 746–750, Singapore, 2014.

90. Paliwal, K. K., "Spectral Subband Centroid Features for Speech Recognition", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 617–620, 1998.

91. Huckvale, M., "Prediction of Cognitive Load from Speech with the VOQAL Voice Quality Toolbox for the INTERSPEECH 2014 Computational Paralinguistics Challenge", *Proceedings of INTERSPEECH*, pp. 741–745, Singapore, 2014.

92. Montacié, C. and M.-J. Caraty, "High-Level Speech Event Analysis for Cognitive Load Classification", *Proceedings of INTERSPEECH*, pp. 731–735, Singapore, 2014.

93. Zhang, Z., E. Coutinho, J. Deng and B. Schuller, "Cooperative Learning and its Application to Emotion Recognition from Speech", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 1, pp. 115–126, 2015.

94. Hotelling, H., "Relations Between Two Sets of Variates", *Biometrika*, Vol. 28, No. 3/4, pp. 321–377, 1936.

95. Hardoon, D. R., S. Szedmak and J. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods", *Neural Computation*, Vol. 16, No. 12, pp. 2639–2664, 2004.

96. Andrew, G., R. Arora, J. Bilmes and K. Livescu, "Deep Canonical Correlation Analysis", *Proceedings of the 30th International Conference on Machine Learning*, pp. 1247–1255, Atlanta, Georgia, USA, 2013.

97. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems", *Annals of Eugenics*, Vol. 7, No. 2, pp. 179–188, 1936.

98. Bartlett, M. S., "Further Aspects of the Theory of Multiple Regression", *Proceedings of the Cambridge Philosophical Society*, Vol. 34, No. 1, pp. 33–40, 1938.

99. Kim, M. and V. Pavlovic, "Central Subspace Dimensionality Reduction using Covariance Operators", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 4, pp. 657–670, 2011.

100. Sakar, C. O., *Multi-View Feature Extraction based on Canonical Correlation Analysis*, Ph.D. Thesis, Boğaziçi University, 2014.

101. Sakar, C. O., O. Kursun and F. Gürgen, "A Feature Selection Method based on Kernel Canonical Correlation Analysis and the Minimum Redundancy Maximum Relevance Filter Method", *Expert Systems with Applications*, Vol. 39, No. 3, pp. 3432–3437, 2012.

102. Nicolaou, M. A., S. Zafeiriou and M. Pantic, "Correlated-spaces Regression for Learning Continuous Emotion Dimensions", *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 773–776, 2013.

103. Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, Boston, 2nd edn., 1990.

104. He, X. and P. Niyogi, "Locality Preserving Projections", S. Thrun, L. Saul and B. Scholkopf (Editors), *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA, 2004.

105. Sugiyama, M., "Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction", *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 905–912, 2006.

106. Huang, G.-B., Q.-Y. Zhu and C.-K. Siew, "Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks", *Proceedings of IEEE International Joint Conference on Neural Networks*, Vol. 2, pp. 985–990, IEEE, 2004.

107. Suykens, J. A. and J. Vandewalle, "Least Squares Support Vector Machine Classifiers", *Neural Processing Letters*, Vol. 9, No. 3, pp. 293–300, 1999.

108. Han, K., D. Yu and I. Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine", *Proceedings of INTERSPEECH*, pp. 223–227, ISCA, Singapore, 2014.

109. Rao, C. R. and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*, Vol. 7, Wiley New York, 1971.

110. Ojala, T., M. Pietikainen and T. Maenpaa, "Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 7, pp. 971–987, 2002.

111. Almaev, T. R. and M. F. Valstar, "Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition", *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII '13)*, pp. 356–361, 2013.

112. Hamm, J. and D. D. Lee, "Grassmann Discriminant Analysis: A Unifying View on Subspace-based Learning", *Proceedings of the 25th International Conference on Machine Learning*, pp. 376–383, 2008.

113. Lovrić, M., M. Min-Oo and E. A. Ruh, "Multivariate Normal Distributions Parametrized as a Riemannian Symmetric Space", *Journal of Multivariate Analysis*, Vol. 74, No. 1, pp. 36–48, 2000.

114. Liu, M., R. Wang, S. Li, S. Shan, Z. Huang and X. Chen, "Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild", *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pp. 494–501, ACM, New York, NY, USA, 2014.

115. Vemulapalli, R., J. K. Pillai and R. Chellappa, "Kernel Learning for Extrinsic Classification of Manifold Features", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 1782–1789, 2013.

116. Arsigny, V., P. Fillard, X. Pennec and N. Ayache, "Geometric Means in a Novel Vector Space Structure on Symmetric Positive-definite Matrices", *SIAM journal on matrix analysis and applications*, Vol. 29, No. 1, pp. 328–347, 2007.

117. Wang, R., H. Guo, L. S. Davis and Q. Dai, "Covariance Discriminative Kearning: A Natural and Efficient Approach to Image Set Classification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2496–2503, 2012.

118. Kursun, O. and E. Alpaydin, "Canonical Correlation Analysis for Multiview Semisupervised Feature Extraction", *Proceedings of the 10th International Conference on Artificial Intelligence and Soft Computing*, pp. 430–436, 2010.

119. Kursun, O., E. Alpaydin and O. V. Favorov, "Canonical Correlation Analysis using Within-Class Coupling", *Pattern Recognition Letters*, Vol. 32, No. 2, pp. 134–144, 2011.

120. Breiman, L., "Random Forests", *Machine learning*, Vol. 45, No. 1, pp. 5–32, 2001.

121. Hall, M. A., *Correlation-Based Feature Selection for Machine Learning*, Ph.D. Thesis, The University of Waikato, 1999.

122. Peng, H., F. Long and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp. 1226–1238, 2005.

123. Pudil, P., J. Novovičová and J. Kittler, "Floating Search Methods in Feature Selection", *Pattern Recognition Letters*, Vol. 15, No. 11, pp. 1119–1125, 1994.

124. Hardoon, D. R., J. Shawe-Taylor and O. Friman, *KCCA Feature Selection for fMRI Analysis*, Technical Report TR_soton_04_03, University of Southampton, 2004.

125. Moearland, P., *Mixture Models for Unsupervised and Supervised Learning*, Ph.D. Thesis, The Swiss Federal Institute of Technology at Lausanne, 2000.

126. McLachlan, G. and D. Peel, *Finite Mixture Models*, New York: Wiley, 2000.

127. Jain, A. K., "Data Clustering: 50 Years Beyond K-means", *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651–666, 2010.

128. Dempster, A. P., N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of The Royal Statistical Society, Series B*, Vol. 39, No. 1, pp. 1–38, 1977.

129. Ghahramani, Z. and G. E. Hinton, *The EM Algorithm for Mixtures of Factor Analyzers*, Technical Report CRG-TR-96-1, University of Toronto, 1997.

130. Salah, A. A. and E. Alpaydın, "Incremental Mixtures of Factor Analysers", *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, Vol. 1, pp. 276–279, 2004.

131. Akaike, H., "A New Look at the Statistical Model Identification", *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 716–723, 1974.

132. Schwarz, G., "Estimating the Dimension of a Model", *Annals of Statistics*, Vol. 6, No. 2, pp. 461–464, 1979.

133. Rissanen, J., "A Universal Prior for Integers and Estimation by MDL", *The Annals of Statistics*, Vol. 11, No. 2, pp. 416–431, 1983.

134. Rissanen, J., *Information and Complexity in Statistical Modeling*, Information Science and Statistics, Springer, Dordrecht, 2007.

135. Wallace, C. and P. Freeman, "Estimation and Inference by Compact Coding", *Journal of Royal Statistical Society, Series B (Methodological)*, Vol. 49, No. 3, pp. 240–265, 1987.

136. Verbeek, J. J., N. Vlassis and B. Kröse, "Efficient Greedy Learning of Gaussian Mixture Models", *Neural computation*, Vol. 15, No. 2, pp. 469–485, 2003.

137. Ghahramani, Z. and M. J. Beal, "Variational Inference for Bayesian Mixtures of Factor Analysers", *Advances in Neural Information Processing Systems*, 2000.

138. Rasmussen, C. E., "The Infinite Gaussian Mixture Model", *Advances in Neural Information Processing Systems*, 11, pp. 554–560, 2000.

139. Gomes, R., M. Welling and P. Perona, "Incremental Learning of Nonparametric Bayesian Mixture Models", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, 2008.

140. Shi, L., S. Tu and L. Xu, "Learning Gaussian Mixture with Automatic Model Selection: A Comparative Study on Three Bayesian Related Approaches", *Frontiers of Electrical and Electronic Engineering in China*, Vol. 6, No. 2, pp. 215–244, 2011.

141. Bouveyron, C. and C. Brunet-Saumard, "Model-Based Clustering of High-Dimensional Data: A Review", *Computational Statistics & Data Analysis*, Vol. 71, pp. 52–78, 2014.

142. Pelleg, D. and A. W. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters", *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 727–734, 2000.

143. Figueiredo, M. A. T. and A. K. Jain, "Unsupervised Learning of Finite Mixture Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 3, pp. 381–396, 2002.

144. Law, M., M. A. T. Figueiredo and A. Jain, "Simultaneous Feature Selection and Clustering Using Mixture Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 9, pp. 1154–1166, 2004.

145. Zivkovic, Z. and F. van der Heijden, "Recursive Unsupervised Learning of Finite Mixture Models", *IEEE Transactions on Pattern Analysis and Machine Intelli-*

gence, Vol. 26, No. 5, pp. 651–656, 2004.

146. Constantinopoulos, C. and A. Likas, "Unsupervised Learning of Gaussian Mixtures Based on Variational Component Splitting", *IEEE Transactions on Neural Networks*, Vol. 18, No. 3, pp. 745–755, 2007.

147. Boutemedjet, S., N. Bouguila and D. Ziou, "A Hybrid Feature Extraction Selection Approach for High-Dimensional Non-Gaussian Data Clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 8, 2009.

148. Gorur, D. and C. E. Rasmussen, "Nonparametric Mixtures of Factor Analyzers", *Proceedings of IEEE Signal Processing and Communications Applications Conference (SIU 2009)*, pp. 708–711, 2009.

149. Yang, M.-S., C.-Y. Lai and C.-Y. Lin, "A Robust EM clustering Algorithm for Gaussian Mixture Models", *Pattern Recognition*, Vol. 45, No. 11, pp. 3950–3961, 2012.

150. Iwata, T., D. Duvenaud and Z. Ghahramani, *Warped Mixtures for Nonparametric Cluster Shapes*, Technical Report arXiv:1206.1846, 2012.

151. Fan, W., N. Bouguila and D. Ziou, "Variational Learning of Finite Dirichlet Mixture Models using Component Splitting", *Neurocomputing*, Vol. 129, pp. 3 – 16, 2014.

152. Fan, W. and N. Bouguila, "Online Variational Learning of Generalized Dirichlet Mixture Models with Feature Selection", *Neurocomputing*, Vol. 126, pp. 166 – 179, 2014.

153. Kersten, J., "Simultaneous Feature Selection and Gaussian Mixture Model Estimation for Supervised Classification Problems", *Pattern Recognition*, Vol. 47, No. 8, pp. 2582 – 2595, 2014.

154. Tipping, M. E. and C. M. Bishop, "Mixtures of Probabilistic Principal Component Analyzers", *Neural Compututation*, Vol. 11, No. 2, pp. 443–482, 1999.

155. Mardia, K. V., J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Probability and Mathematical Statistics, Academic Press London, 1979.

156. Celeux, G., S. Chrétien, F. Forbes and A. Mkhadri, "A Component-Wise EM Algorithm for Mixtures", *Journal of Computational and Graphical Statistics*, Vol. 10, No. 4, pp. 697–712, 2001.

157. Ueda, N., R. Nakano, Z. Ghahramani and G. E. Hinton, "SMEM Algorithm for Mixture Models", *Neural Computation*, Vol. 12, No. 9, pp. 2109–2128, 2000.

158. Vinh, N. X., J. Epps and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance", *J. Mach. Learn. Res.*, Vol. 11, pp. 2837–2854, 2010.

159. Geva, A. B., "Hierarchical Unsupervised Fuzzy Clustering", *IEEE Transactions on Fuzzy Systems*, Vol. 7, No. 6, pp. 723–733, 1999.

160. Samaria, F. S. and A. C. Harter, "Parameterisation of a Stochastic Model for Human Face Identification", *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*, pp. 138–142, 1994.

161. Kohonen, T., J. Hynninen, J. Kangas and K. Torkkola, "LVQ-PAK", Helsinki University of Technology, 1995.

162. Gürgen, F. S., R. Alpaydin, U. Ünlüakin and E. Alpaydin, "Distributed and Local Neural Classifiers for Phoneme Recognition", *Pattern Recognition Letters*, Vol. 15, No. 10, pp. 1111–1118, 1994.

163. LeCun, Y., L. Bottou, Y. Bengio and P. Haffner, "Gradient-based Learning Applied to Document Recognition", *Proceedings of the IEEE*, Vol. 86, No. 11, pp.

2278–2324, 1998.

164. Frank, A. and A. Asuncion, *UCI Machine Learning Repository*, `http://archive.ics.uci.edu/ml`, [Accessed May 2015].

165. Lyakso, E., O. Frolova, E. Dmitrieva, A. Grigorev, H. Kaya and A. A. Karpov, "EmoChildRu: Emotional Child Russian Speech Corpus ", *(submitted)*, 2015.

166. Lyakso, E. E., O. V. Frolova, A. V. Kurazhova and J. S. Gaikova, "Russian Infants and Children's Sounds and Speech Corpuses for Language Acquisition Studies", *Proceedings of INTERSPEECH*, pp. 1888–1891, 2010.

167. Eyben, F., F. Weninger, F. Groß and B. Schuller, "Recent Developments in openSMILE, the Munich open-source Multimedia Feature Extractor", *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 835–838, ACM, 2013.

168. Schuller, B., S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller and S. S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge.", *Proceedings of INTERSPEECH*, pp. 2794–2797, 2010.

169. Schuller, B., S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition", *Proceedings of INTERSPEECH*, Dresden, Germany, 2015.

170. Bach, F. R. and M. I. Jordan, *A Probabilistic Interpretation of Canonical Correlation Analysis*, Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

171. Cai, Z., L. Wang, X. Peng and Y. Qiao, "Multi-view Super Vector for Action Recognition", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 596–603, 2014.

172. Moore, J., L. Tian and C. Lai, "Word-Level Emotion Recognition Using High-Level Features", A. Gelbukh (Editor), *Computational Linguistics and Intelligent Text Processing*, Vol. 8404 of *Lecture Notes in Computer Science*, pp. 17–31, Springer Berlin Heidelberg, 2014.

173. Kim, S., M. Filippone, F. Valente and A. Vinciarelli, "Predicting Continuous Conflict Perception with Bayesian Gaussian Processes", *IEEE Transactions on Affective Computing*, Vol. 5, No. 2, pp. 187–200, 2014.

174. Torres, J., A. Saad and E. Moore, "Evaluation of Objective Features for Classification of Clinical Depression in Speech by Genetic Programming", *Proceedings of 11th World Conference on Soft Computing in Industrial Applications*, Vol. 39, pp. 132–143, Springer Berlin Heidelberg, 2007.

175. Park, C.-H. and K.-B. Sim, "The Novel Feature Selection Method Based on Emotion Recognition System", D.-S. Huang, K. Li and G. W. Irwin (Editors), *Computational Intelligence and Bioinformatics*, Vol. 4115 of *Lecture Notes in Computer Science*, pp. 731–740, Springer Berlin Heidelberg, 2006.

176. Torres, J., A. Saad and E. Moore, "Application of a GA/Bayesian Filter-Wrapper Feature Selection Method to Classification of Clinical Depression from Speech Data", A. Saad, K. Dahal, M. Sarfraz and R. Roy (Editors), *Proceedings of 11th World Conference on Soft Computing in Industrial Applications*, Vol. 39, pp. 115–121, Springer Berlin Heidelberg, 2007.

177. Espinosa, H., J. Garcia and L. Pineda, "Bilingual Acoustic Feature Selection for Emotion Estimation using a 3D Continuous Model", *Proceedings of the 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pp. 786–791, 2011.

178. Giannoulis, P. and G. Potamianos, "A Hierarchical Approach with Feature Selection for Emotion Recognition from Speech", N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and

S. Piperidis (Editors), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, 2012.

179. Kirchhoff, K., Y. Liu and J. Bilmes, "Classification of Developmental Disorders from Speech Signals using Submodular Feature Selection", *Proceedings of INTERSPEECH*, pp. 187–190, ISCA, Lyon, France, 2013.

180. Bejani, M., D. Gharavian and N. M. Charkari, "Audiovisual Emotion Recognition using ANOVA Feature Selection Method and Multi-classifier Neural Networks", *Neural Computing and Applications*, Vol. 24, No. 2, pp. 399–412, 2014.

181. Kim, S., M. Filippone, F. Valente and A. Vinciarelli, "Predicting the Conflict Level in Television Political Debates: An Approach based on Crowdsourcing, Nonverbal Communication and Gaussian Processes", *Proceedings of the 20th ACM International Conference on Multimedia*, pp. 793–796, 2012.

182. Vinciarelli, A., S. Kim, F. Valente and H. Salamin, "Collecting Data for Socially Intelligent Surveillance and Monitoring Approaches: The Case of Conflict in Competitive Conversations", *International Symposium on Communications, Control, and Signal Processing*, pp. 1–4, 2012.

183. Platt, J., *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*, Technical Report MSR-TR-98-14, Microsoft Research, 1998.

184. Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update", *ACM SIGKDD Explorations Newsletter*, Vol. 11, No. 1, pp. 10–18, 2009.

185. Harada, S., J. Lester, K. Patel, T. S. Saponas, J. Fogarty, J. A. Landay and J. O. Wobbrock, "VoiceLabel: Using Speech to Label Mobile Sensor Data", *Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI '08)*, pp.

69–76, 2008.

186. Schuller, B., F. Friedmann and F. Eyben, "The Munich Bio Voice Corpus: Effects of Physical Exercising, Heart Rate, and Skin Conductance on Human Speech Production", *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 2014.

187. Dehak, N., P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end Factor Analysis for Speaker Verification", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798, 2011.

188. Kenny, P., G. Boulianne, P. Ouellet and P. Dumouchel, "Joint Factor Analysis versus Eigenchannels in Speaker Recognition", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 4, pp. 1435–1447, 2007.

189. Sivic, J. and A. Zisserman, "Efficient Visual Search of Videos Cast as Text Retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 4, pp. 591–606, 2009.

190. Wang, W., P. Lu and Y. Yan, "An Improved Hierarchical Speaker Clustering", *Acta Acustica*, Vol. 33, No. 1, p. 9, 2008.

191. Han, K. J., S. Kim and S. S. Narayanan, "Strategies to Improve the Robustness of Agglomerative Hierarchical Clustering Under Data Source Variation for Speaker Diarization", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 8, pp. 1590–1601, 2008.

192. Varol, G. and A. A. Salah, "Extreme Learning Machine for Large-scale Action Recognition", *ECCV Workshop on Action Recognition with a Large Number of Classes*, 2014.

193. Perronnin, F., J. Sánchez and T. Mensink, "Improving the Fisher Kernel for Large-scale Image Classification", *Proceedings of the 11th European Conference*

on Computer Vision, pp. 143–156, 2010.

194. Hantke, S., F. Weninger, R. Kurle, A. Batliner and B. Schuller, "I Hear You Eat and Speak: Automatic Recognition of Eating Condition and Food Type", *(to appear)*.

195. Ellis, D. P. W., *PLP and RASTA (and MFCC, and inversion) in Matlab*, 2005, `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`, [Accessed May 2015].

196. Vedaldi, A. and B. Fulkerson, *VLFeat: An Open and Portable Library of Computer Vision Algorithms*, 2008, `http://www.vlfeat.org/`, [Accessed May 2015].

197. MATLAB, *Statistical Toolbox version 9.0 (R2014a)*, The MathWorks Inc., Natick, Massachusetts, 2014.

198. Dhall, A., R. Goecke, J. Joshi, M. Wagner and T. Gedeon, "Emotion Recognition in the Wild Challenge 2013", *Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI '13)*, pp. 509–516, 2013.

199. Liu, M., R. Wang, Z. Huang, S. Shan and X. Chen, "Partial Least Squares Regression on Grassmannian Manifold for Emotion Recognition", *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pp. 525–530, 2013.

200. Sun, B., L. Li, T. Zuo, Y. Chen, G. Zhou and X. Wu, "Combining Multimodal Features with Hierarchical Classifier Fusion for Emotion Recognition in the Wild", *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pp. 481–486, ACM, New York, NY, USA, 2014.

201. Lowe, D. G., "Distinctive Image Features from Scale-invariant Keypoints", *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110, 2004.

202. McNemar, Q., "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages", *Psychometrika*, Vol. 12, No. 2, pp. 153–157, 1947.

203. Kahou, S. E., C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda and Z. Wu, "Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video", *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI '13)*, pp. 543–550, 2013.

204. Shan, C., S. Gong and P. W. McOwan, "Capturing Correlations Among Facial Parts for Facial Expression Analysis", *Proceedings of British Machine Vision Conference*, pp. 1–10, 2007.

205. Shan, C., S. Gong and P. W. McOwan, "Beyond Facial Expressions: Learning Human Emotion from Body Gestures", *Proceedings of British Machine Vision Conference*, pp. 1–10, 2007.

206. Sánchez-Lozano, E., P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rúa and J. L. Alba-Castro, "Audiovisual Three-level Fusion for Continuous Estimation of Russell's Emotion Circumplex", *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '13)*, pp. 31–40, 2013.

207. Beck, A., R. Steer, R. Ball and W. Ranieri, "Comparison of Beck Depression Inventories -IA and -II in Psychiatric Outpatients", *Journal of Personality Assessment*, Vol. 67, No. 3, pp. 588–597, 1996.

208. Eyben, F., F. Weninger, S. Squartini and B. Schuller, "Real-life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pp. 483–487, 2013.

209. Jiang, B., M. F. Valstar and M. Pantic, "Action Unit Detection Using Sparse Appearance Descriptors in Space-time Video Volumes", *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011)*, pp. 314–321, 2011.

210. Xiong, X. and F. De la Torre, "Supervised Descent Method and Its Application to Face Alignment", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 532–539, 2013.

211. Sakar, C. O., O. Kursun and S. F. Gurgen, "Ensemble Canonical Correlation Analysis", *Applied Intelligence*, Vol. 40, No. 2, pp. 291–304, 2014.

212. Cummins, N., J. Joshi, A. Dhall, V. Sethu, R. Goecke and J. Epps, "Diagnosis of Depression by Behavioural Signals: A Multimodal Approach", *Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge (AVEC '13)*, pp. 11–20, 2013.

213. Meng, H., D. Huang, H. Wang, H. Yang, M. AI-Shuraifi and Y. Wang, "Depression Recognition Based on Dynamic Facial and Vocal Expression Features Using Partial Least Square Regression", *Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge (AVEC '13)*, pp. 21–30, 2013.

214. Cummins, N., J. Epps, V. Sethu and J. Krajewski, "Variability Compensation in Small Data: Oversampled Extraction of i-Vectors for the Classification of Depressed Speech", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, pp. 970–974, Florence, Italy, 2014.

215. Kaya, H., "Speaker- and Corpus-Independent Methods for Affect Classification in Computational Paralinguistics", *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pp. 359–363, ACM, 2014.