Affective Speech Recognition

by

Seyyed Pouria Fewzee Youssefi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2015

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Speech, as a medium of interaction, carries two different streams of information. Whereas one stream carries explicit messages, the other one contains implicit information about speakers themselves. Affective speech recognition is a set of theories and tools that intend to automate unfolding the part of the implicit stream that has to do with humans emotion. Application of affective speech recognition is to human computer interaction; a machine that is able to recognize humans emotion could engage the user in a more effective interaction. This thesis proposes a set of analyses and methodologies that advance automatic recognition of affect from speech. The proposed solution spans two dimensions of the problem: speech signal processing, and statistical learning.

At the speech signal processing dimension, extraction of speech low-level descriptors is discussed, and a set of descriptors that exploit the spectrum of the signal are proposed, which have shown to be particularly practical for capturing affective qualities of speech. Moreover, considering the non-stationary property of the speech signal, further proposed is a measure of dynamicity that captures that property of speech by quantifying changes of the signal over time. Furthermore, based on the proposed set of low-level descriptors, it is shown that individual human beings are different in conveying emotions, and that parts of the spectrum that hold the affective information are different from one person to another. Therefore, the concept of emotion profile is proposed that formalizes those differences by taking into account different factors such as cultural and gender-specific differences, as well as those distinctions that have to do with individual human beings.

At the statistical learning dimension, variable selection is performed to identify speech features that are most imperative to extracting affective information. In doing so, low-level descriptors are distinguished from statistical functionals, therefore, effectiveness of each of the two are studied dependently and independently. The major importance of variable selection as a standalone component of a solution is to real-time application of affective speech recognition. Although thousands of speech features are commonly used to tackle this problem in theory, extracting that many features in a real-time manner is unrealistic, especially for mobile applications. Results of the conducted investigations show that the required number of speech features is far less than the number that is commonly used in the literature of the problem.

At the core of an affective speech recognition solution is a statistical model that uses speech features to recognize emotions. Such a model comes with a set of parameters that are estimated through a learning process. Proposed in this thesis is a learning algorithm, developed based on the notion of Hilbert-Schmidt independence criterion and named max-dependence regression, that maximizes the dependence between predicted and actual values of affective qualities. Pearson's correlation coefficient is commonly used as the measure of goodness of a fit in the literature of affective computing, therefore max-dependence regression is proposed to make the learning and hypothesis testing criteria consistent with one another. Results of this research show that doing so results in higher prediction accuracy.

Lastly, sparse representation for affective speech datasets is considered in this thesis. For this purpose, the application of a dictionary learning algorithm based on Hilbert-Schmidt independence criterion is proposed. Dictionary learning is used to identify the most important bases of the data in order to improve the generalization capability of the proposed solution to affective speech recognition. Based on the dictionary learning approach of choice, fusion of feature vectors is proposed. It is shown that sparse representation leads to higher generalization capability for affective speech recognition.

Acknowledgements

I acknowledge that the realization of this thesis wouldn't have been possible if it wasn't for the support of my PhD supervisor Dr. Fakhri Karray.

I am thankful to Dr. Abdoulmotaleb El Saddik, Dr. Mohamed Kamel, Dr. Dana Kulić, Dr. John Zelek, and Dr. Ali Ghodsi, as the committee members of my comprehensive exam and final defense sessions.

Dedication

I dedicate this thesis to my family, my friends, and to those who direct their everyday efforts towards the realization of their present happiness, as well as that of their fellow beings.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AIB Agglomerative Information Bottleneck

AIC Akaike Information Criterion

AVEC Audio/Visual Emotion Challenge

CC Correlation Coefficient

CFA Cross-modal Factor Analysis

CV Cross Validation

DFT Discrete Fourier Transform

DLSI Dictionary Learning with Structured Incoherence

DLSR Dictionary Learning and Sparse Representation

ELNET Elastic Net

FCSC Fully-Continuous Sub-Challenge

FDDL Fisher Discrimination Dictionary Learning

FS Fisher Score

HCI Human Computer Interaction

HMM Hidden Markov Model

HNR Harmonic-to-Noise Ratio

HSIC Hilbert-Schmidt Independence Criterion

KLD Kullback-Leibler Divergence

LLD Low-Level Descriptor

LOSO Leave-One-Subject-Out

LPC Linear Predictive Coefficient

LS Least Squares

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MDR Max-Dependence Regression

MFCC Mel-Frequency Cepstrum Coefficient

MLE Maximum Likelihood Estimator

MSE Mean Squared Error

MV Multi-View

MVSDL Multi-View Supervised Dictionary Learning

NF Numer of Features

NNZ Number of Non-Zero values

OCEAN Openness-Conscientiousness-Extraversion-Agreeableness-Neuroticism

OLS Ordinary Least Squares

PAD Pleasure-Arousal-Dominance

PCA Principle Component Analysis

RBF Radial Basis Function

RCF Randomized Clustering Forests

RF Random Forests

RGLM Random Generalized Linear Models

RKHS Reproducing Kernel Hilbert Space

RMS Root Mean Squared

RMSE Root Mean Squared Error

SC Sub-Challenge

SDL Supervised Dictionary Learning

SED Spectral Energy Distribution

SEP Spectral Emotion Profile

SLD Speaker Likability Database

SPC Supervised Principal Components

SPCA Supervised Principal Component Analysis

SRC Sparse Representation-based Classification

SVD Singular Value Decomposition

SVM Support Vector Machines

SVR Support Vector Regression

UA Unweighted Average

UAR Unweighted Average Recall

WA Weighted Average

WLSC Word-Level Sub-Challenge

ZCR Zero-Crossing Rate

Chapter 1

Introduction

Understanding affect in human behavior has been a topic of interest to researchers from different disciplines for many years. It has early roots in Charles Darwin's works [38] and it is brought to much more maturity by social and behavioral psychologists, as well as cognitive scientists [95]. More recently, by introducing affective computing, the research of affection has also involved computer science and engineering researchers. Affective computing, as Picard defines [151], is *"computing that relates to, arises from, or deliberately influences emotion."*

A natural application of affective computing is to human-computer interaction (HCI). That is, to enable computers to adapt to emotional states of users in order to reduce their frustration during interactions [150]. Different modalities (also referred to as social cues) have been used for this purpose, among which only vocal cues have led to the current research. Although automatic speech recognition has been around for many years, often times one would like to go beyond the point of knowing *what* is said in a conversation, and one would like to understand *how* they are said. Speech acts convey much more information than mere literal verbal content [24] and affective speech recognition comes to reveal those information.

In this chapter, the problem of affective speech recognition is stated, as concerns the mathematical modeling side of the problem. Next, the scope of the dissertation is discussed by enumerating some of the difficulties associated with the problem, as well as proposed solutions to those problems. This chapter is concluded by outlining the dissertation.

1.1 Problem Statement

Given a speech signal s, affective speech recognition is aimed at predicting the emotional content of s, represented by $\mathbf{y}$. To do so, the problem is handled in two stages. The first stage, which falls under the focus of speech signal processing, deals with identification and extraction of speech features. The output of this stage is a set of speech features $\mathbf{x}$.

$$g : s \rightarrow \mathbf{x} \tag{1.1}$$

The second stage, on the other hand, falls under the focus of statistical learning. The output of this stage is a model f that describes the interrelation between $\mathbf{x}$ and $\mathbf{y}$.

$$f : \mathbf{x} \rightarrow \mathbf{y} \tag{1.2}$$

The objective of affective speech recognition is hence to define g and f, so that the overall system, i.e., $f(g(s))$, can approximate the emotional content of an utterance.

1.2 Challenging Aspects

To address some challenges of affective speech recognition, the following questions are asked:

1. *What set of speech features can carry its affective properties?* On the one hand, thousands of speech features are commonly extracted for speech signal processing, however, not all of those features are suitable for capturing affective contents of speech. On the other hand, the existing set of features are not guaranteed to be sufficient for that purpose. Therefore, a proper solution to this problem may suggest a concise subset of existing, and possibly new, speech features that can serve the purpose. Furthermore, since speech features are composed of low-level descriptors and statistical functionals, the solution should address these details as well. That is, what set of low-level descriptors, and what set of statistical functionals are suitable for capturing affective contents of speech.

2. *Does an answer to the first question depend on individual human beings?* It is claimed that expression of emotions are person-specific [95]. That is, individual human beings, as well as their supersets that share some characteristics, e.g., gender, are different in conveying affect. Similarly, cultural background of speakers can contribute to the way they express affect [137, 192]. Therefore, the question is whether these sources of variation can have an impact on a selected set of speech features. This information could be useful for

defining individual profiles of affect, as well as those of groups of individuals. That is, to customize the set of selected speech features that take into account those varying factors. Such emotion profiles could enable personalization in different levels.

3. *What statistical learning algorithms are suitable for estimating parameters of a model as described in Equation 1.2?* In spite of the fact that at the core of any learning algorithm there is the component of estimating some parameters of some sort, different applications require different optimization criteria. Therefore, an estimate, although optimal from one perspective, may not be optimal from another. Therefore, the question is whether the existing learning algorithms optimize what is desired to be optimized, and if not, how could one improve on that.

4. *In the presence of multitude of learning patterns, what combination of those can concisely capture their essence?* Oftentimes, there is an abundance of affective speech samples for training a model. Therefore, finding a handful that can sufficiently and briefly describe the key variations becomes challenging. A representation as such, if available, would promise higher generalization capabilities, and lower computational expenses. Therefore, we would like to investigate if such a representation exists for affective speech, and if so, how many bases would suffice to meet those objectives.

Our contributions to affective speech recognition are attempts to answering these questions.

1.3 Proposed Solutions and Contributions

To answer the *first* question, we start by proposing a set of low-level descriptors that take advantage of the information that's spread along the spectrum, which we call spectral energy distribution or SED. In addition, based on SED, we define a measure of dynamicity to quantify temporal changes of the speech signal.

Furthermore, to understand what set of speech features are useful for modeling affect, we run several dimensionality reduction and variable selection experiments. As a result of those experiments, we show what set of features are useful for capturing affective contents of speech, although our experiences are specific to our choice of dataset, and that may or may not be generalized all sorts of scenarios. In doing so, we emphasize on the length of the selected set of features, and show that, for some affective dimensions, very few number of features suffice to explain the major part of the variation, whereas for some others, a considerably longer set of features is required. We also show that less number of features results in higher generalization capability in most of cases. In discussing a selected set of features, we distinguish the difference between low-level descriptors and statistical functionals, and we show that particular statistical functionals are useful for certain low-level descriptors. By doing so, we encourage selective feature extraction for affective speech recognition, as opposed to using a long list of statistical functionals for the contours of a long list of low-level descriptors, i.e., brute force extraction.

Our proposed answer to the *second* question is based on the definition of SED. We introduce the concept of spectral emotion profile (SEP) to formalize individuals' differences in conveying affect that are reflected on the spectrum of their speech. Using SEP, we verify that the optimal set of speech features vary dramatically from one individual speaker to another, as well as from one gender to the other. That is to say, according to the definition of spectral energy distribution, the spectral intervals that are efficient for capturing affect are different from one person to another, and that the choice of those intervals may also depend on the gender of a speaker.

In the literature of continuous recognition of affect, Pearson's correlation coefficient is frequently used as a measure of goodness of fit. Therefore, in order to answer the *third* question, we introduce max-dependence regression (MDR). To do so, we make use of the Hilbert-Schmidt independence criterion as a generic measure of independence. Unlike the existing learning algorithms, that minimize a sense of prediction error, MDR's estimate for the coefficients of the linear model are those that maximize dependency of the predicted response variable on their actual values. By doing so, MDR synchronizes the optimization criterion with the hypothesis testing criterion that is commonly used for assessing the goodness of a fit in affective computing.

In order to answer the *forth* question, we introduce dictionary learning to affective speech recognition. By defining dictionaries over affective speech, we summarize affective speech data into a set of atoms that serve as bases of the data, in the sense that data samples can be reconstructed using linear combinations of those bases. Using dictionary learning, we further propose fusion of different speech features at different levels.

1.4 Organization

This dissertation is organized as follows. In Chapter 2 we go over the literature of affective speech recognition. To do so, we review the conventional set of speech features that are commonly extracted for modeling affective speech. Moreover, we review different statistical modeling approaches that are found in the literature of the problem, in terms of dimensionality reduction, classification, and regression models and algorithms that are used for affective speech recognition. Finally, we review the affective speech datasets that have been served as the framework of numerous studies.

In Chapter 3 we highlight theoretical background of the research. We start by reviewing the conventional ways of representing affect. Then, we review the common procedure for extracting features from speech, and we review the definition of most common of those features that we use in this thesis. Finally, we end this chapter by going over the statistical learning approaches that we use throughout the research.

Chapter 4 presents our contributions to the speech signal processing part of the problem. We start by introducing spectral energy distribution (SED) as a set of speech features for affective speech recognition. Then, based on SED, we propose a measure of dynamicity that tries to quantify temporal variations of the speech signal. Finally, in this chapter, we introduce the concept of emotional speech profile (SEP).

Chapter 5 captures our contributions to the statistical learning side of affective speech recognition. We start by discussing variable selection for paralinguistic speech recognition. Then, we introduce max-dependence regression for estimating the parameters of the linear model. Finally, we introduce dictionary learning to affective speech recognition.

Chapter 6 summarizes contributions of the research for dealing with major challenges outlined earlier in the thesis.

Chapter 2

Literature Review

In this chapter, we review the literature of affective speech recognition. We go over the literature of extraction of acoustic features from speech. Then, we highlight different methods that are used for modeling affective speech. And finally, we review the most well-known and widely used affective speech datasets.

2.1 Acoustic Features

In this Section we review the prevalent set of speech features that are used in the literature of affective speech recognition, as well as those that have recently appeared in the literature of the problem. Then, we metion their shortcomings, and explain how those shortcomings are addressed in this thesis.

Speech features are compounds of low-level descriptors and statstical functionals. Tabels 2.1 and 2.2 list the most commonly used low-level descriptors and statistical functionals in the literature of affective speech recognition.

Besides the listed set of features, a set of 18 features known as VOQAL features developed at the University College London were used by Hu [87]; Asgari *et al.* [7] used fundamental frequency, jitter, shimmer, and harmonic-to-noise ratio; Bone *et al.* [17] made use of a set of different speech features including pitch, duration of phonemes, formants, intensity, goodness of prononciation, and spectral energy; Martinez *et al.* [122] used prosodic features, formant modeling, Legendre polynomials, shifted-delta cepstral coefficients, amplitude modulation index, and speaking rate; Oh *et al.* [145] used syllabic-level segmentation and extracted features such as

Table 2.1: Listing of the types of audio features in the feature vector [188]

Energy & spectral (25)
loudness (auditory model based), zero crossing rate, energy in bands from 250-650 Hz, 1-4 kHz, 25%, 50%, 75%, and 90% spectral roll-off points, spectral flux, entropy, variance, skewness, kurtosis, psychoacousitc sharpness, harmonicity, MFCC 1-10
Voicing related (6)
F_0 (sub-harmonic summation, followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: "jitter of jitter"), logarithmic Harmonics-to-Noise Ratio (logHNR)

Table 2.2: Listing of the statistics computed for audio features. ([1]: Not applied to delta coefficient contours. [2]: For delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied. [3]: Not applied to voicing related LLD.) [188]

Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1%, 99% percentile, percentile range 1%-99%, percentage of frames contour is above: minimum + 25%, 50%, and 90% of the range, percentage of frames contour is rising, maximum, mean, minimum segment length[1,3], standard deviation of segment length[1,3]
Regression functionals[1] (4)
linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient a, and approximation error (linear)
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude mean of minima, amplitude range of maxima
Other[1,3] (6)
LP gain, LPC 1-5

intensity, pitch, timbre, and rhythmic patterns; Trancoso *et al.* [5] extracted text-derived features such as cepstral distortion, speaking rate, and statistics of phoneme duration; Buisman and Postma [20] used features based on Gabor filter, which roots in spectrogram analysis in image preocessing; Cummins *et al.* [36] used pitch, pitch direction, jitter, shimmer, zero-corssing rate, spectral roll-off, and spectral flux, as well as cepstral coefficients, line spectral pairs, and spectral central frequencies and amplitudes; Kim *et al.* [96] used multiple language phoneme probability features, as well as prosodic and informational features, in addition to voice quality and prononciation features; Montacié and Caraty [130] used pitch and intonation features; Sanchez *et al.* [169] extracted basic prosodic features, prosodic polynomial coefficients, mel frequency cepstral coefficients, and shifted delta cepstrum; Zhou *et al.* [241] used spectro-temporal features; Amarakeerthi *et al.* [2] extracted two-layered cascaded subband filter and cepstral coefficients; Nomoto *et al.* [141] used linguistic features; Kitahara *et al.* [100] extracted segment duration ratio features; Luengo *et al.* [117] used low-frequency power coefficients, as well as intonation, power, rhythm, voice quality, and sentence-end feautres; Bozkurt *et al.* [18] extracted dynamic, as well as HMM-based features; and Polzehl *et al.* [157] used glottal excitation features.

Considering that the generation of speech depends on the vibration of the vocal folds, spectrum of the speech signal should offer considerable amount of relevant information. However, other than the two intervals 250-650 Hz and 1-4 KHz, the literature of affective speech recognition before this research does not show taking advantage of the information that is provided by the spectrum of the signal. Therefore, as a part of this research, in Chapter 4.1 we focus on the development and assessment of a set of speech features that break the spectrum down to fine intervals in order to benefit from every part of the spectral domain that might be found beneficial at the statistical learning stage.

Furthermore, in order to take account of the dynamic nature of the speech signal by analysing it over time, the literature of the probelm suggests using the first difference contour of the low-level descriptors, as well as calculating linear prediction coefficients. The literature of this problem does not show comparing the spectrum of the signal over time, in order to quantify the pace of changes over time. Therefore, also proposed in this research (Chapter 4.2) is a measure of dynamicity that suggests comparative analysis of the spectrum of the speech signal over time.

2.2 Statistical Modeling

Various statistical models are used for dimensionality reduction, classification, and regression purposes for affective speech recognition. In the following, we briefly review those models and algorithms from each of the three categories that are exploited in the most recent literature of affective speech recognition.

2.2.1 Dimensionality Reduction

A wide range of dimensionality reduction algorithms are used in the literature of affective speech recognition, among which filter methods are found most popular. Some examples of such methods are correlation based feature selection, which have been used in studies such as Casale *et al.* [27], Espinosa *et al.* [46], Vogt [211], and Wöllmer *et al.* [218]; or information theoretic selection algorithms, which have been used in various studies including those by Busso *et al.* [23], Dongrui Wu [224], and Chung-Hsien Wu and Wei-Bin Liang [222]; as well as Fisher score that is used in some studies such as those by Bone *et al.* [17], Nomoto *et al.* [141], and Siqing Wu *et al.* [227].

As examples of wrapper methods, forward selection and backward elimination are used by Chastagnol and Devillers [31], Schuller *et al.* [174, 182, 183], Wu *et al.* [226, 227], and Yeh and Chi [233]; also, genetic algorithms are used for this purpose in works by Morrison *et al.* [133], and Wu [224].

Different projection methods have also been adopted in various studies, among which principal component analysis is utilized in studies by Calix *et al.* [25], Espinosa *et al.* [46], Fewzee and Karray [53], Rong *et al.* [166], and Stark *et al.* [197]; and Fisher discriminant analysis and supervised principal component are supervised projection methods that are used in studies by Siqing Wu *et al.* [227] and Fewzee and Karray [53], respectively.

Despite the use of wide range of dimensionality reduction algorithms, the literature of this problem misses in-detail analyses of the selected features. That is, the literature of the problem takes advantage of lower number of features in order to gain lower complexity and higher generalization capabilities, however, the choice of the selected features is left out. That is an import part of feature selection, given that doing such analysis would let use gain a deeper understanding of the relevant set of low-level descriptors and statistical functional that could explain the variation caused by affective contents of speech. On the other hand, by identifying useful sets of features, and by ruling out those that are irrelevant, we could save a considerable amount of computational expences for extraction of features, as well as for parameter estimation for the statistical modeling. Dimensionality reduction and variable selection is the topic of Chapter 5.1.

Furthermore, it is shown that the expression of affect differs from one person to another, and that it is highly correlated with the cultural backgrounds of the speakers [137, 192]. However, the literature of this problem does not show any adaptation to this fact at the feature level, and leaves this distinction to be done at the learning stage where the parameters of the statistical model are estimated. However, such information, e.g., demographics of speakers, could be used in order to direct variable selection towards selecting features that explain variations that are caused by such contexts. Therefore, a part of this research has been devoted to the proposition of the notion of spectral emotion profile, that is introduced in order to examine the spectrum of the speech signal in order to understand which parts of the spectrum are useful for extracting affective contents of speech for different individuals as well as different genders. Spectral emotion profile is the focus of the Chapter 4.3.

2.2.2 Classification

Classification is used in the literature of affective speech recognition, when the categorical model of affect is considered. Support vector machines is one of the most commonly used classifiers in the literature of affective speech recognition, and it is used in various studies such as those by Bone *et al.* [17], Espinosa *et al.* [149], Eyben *et al.* [49], Pierre-Yves [152], Schuller and Devillers [181], and Schuller *et al.* [176, 179, 185, 187].

Gaussian mixture models are as well among the most common classifiers in the literature, where some of the works who have utilized it are as follows: Busso *et al.* [23], Janicki [90], Kockmann *et al.* [101, 102], Dumouchel *et al.* [41], Wu and Liang [222], and Zhou *et al.* [241].

K-nearest neighbors is as well used for recognition of affective speech in works including those by Bone *et al.* [17], Pierre-Yves [152], and Yacoub *et al.* [229].

Decision trees are also used for modeling affect by Burkhardt *et al.* [22], Rong *et al.* [166], Schuller *et al.* [174, 183], and Ślot *et al.* [191].

Furthermore, various types of neural networks are used in different studies including those by Morrison *et al.* [133], Polzehl *et al.* [157], and Wu and Liang [222], who used multi-layer perceptron; Brueckner [19] and Lee *et al.* [111], who used deep neural networks; Brueckner [19], who made use of Boltzmann machines; and Caridakis *et al.* [26], who used recurrent neural networks.

Additionally, different types of probabilistic models are used for the recognition of affect, among which naïve Bayes is utilized by Casale *et al.* [27] and Vogt and André [211]; Bayesian networks are used in studies including those by Fersini *et al.* [52], Kim *et al.* [96], and Schuller *et al.* [174, 175]; studies such as Nwe *et al.* [142], Soleymani *et al.* [194], and Yu *et al.* [234] used hidden Markov models; logistic regression was used in some other works including those by Busso *et al.* [23], Kockmann *et al.* [101], Lee *et al.* [110], Meinedo and Trancoso [127], and Dumouchel *et al.* [41]; linear discriminant analysis is applied in works by Laukka *et al.* [107], Neiberg and Gustafson [136], and Vankayalapati *et al.* [206]; Gaussian processes were used by Lu [116], and Neyman-Pearson lemma by Vlasenko *et al.* [210].

Ensemble learning methods have also been used in works such as Morrison *et al.* [133], which have used random forests, and others including Gosztolya *et al.* [70] and Ivanov and Chen [88], which have use AdaBoost.

Finally, some methods such as sparse representation classification and fuzzy systems were used in studies by Cummins *et al.* [36] and Grimm *et al.* [76], respectively.

Although classification algorithms comprise a significant portion of the literature of affective speech recognition, this research does not focus on classification, since it is used for the recognition of categorical affect, which seems to be going to be fully replaced by the dimensional notion of affect, given that the latter notion subsumes the definition of the former one. In this work, where we focus on the categorical notion of emotion, we use either support vector machines or linear classifier with the Gaussian error assumption.

2.2.3 Regression

A wide range of models and learning algorithms are exploited for statistical modeling of continuous affect, among which support vector regression is found most popular, and is used in studies including those by Espinosa *et al.* [47], Eyben *et al.* [48], Kanluan *et al.* [92], Nicolaou *et al.* [139], Schuller *et al.* [188], Wöllmer *et al.* [218], Siqing Wu *et al.* [226], and Dongrui Wu *et al.* [223, 225].

Moreover, various types of neural networks are utilized for that purpose in works by Caridakis *et al.* [26] and Wöllmer *et al.* [218] that used recurrent neural networks, Gunes *et al.* [78], Nicolaou *et al.* [139], Wöllmer *et al.* [218, 220], who used long short-term memory, Gosztolya *et al.* [70], who used deep neural networks, and Cen *et al.* [28] that use multi-layer perceptron.

Additionally, shrinkage models like lasso are used in different studies including those by Cen *et al.* [28] and van der Maaten [204]. And, Markov models have been also used for continuous recognition of affect by Ozkan *et al.* [146] and Glodek *et al.* [68]. Furthermore, fuzzy inference systems are explored in studies such as those by Soladié *et al.* [193] and Grimm *et al.* [77]. In addition, other types of regression techniques such as Gaussian mixture models by Glodek *et al.* [69], decision trees by Burkhardt *et al.* [22], Gaussian process regression by Kim *et al.* [98], kernel regression by Nicolle *et al.* [140], and autoregressive modeling by Savran *et al.* [170] are seen in the literature.

Closely studying each of the utilised algorithms for estimating the parameters of the regression model, it is noticed that despite the wide range of algorithms that seen in the literature of the problem, what all these different algorithms share is the fact that they optimize for lower prediction error. On the other hand, Pearson's correlation coefficient has been adopted in the literature of affective computing as the standard measure of goodness of fit. That is, among two models, the one which results in a higher correlation coefficient would be favored over the other one. Therefore, a question that we have asked in this research is whether minimizing the prediction error has the same meaning as maximizing the correlation coefficient. Upon having shown that this may not be the case, it is the then proposed in this work a learning algorithm for regression problem that is based on maximizing the correlation coefficient. For this purpose, Hilbert-Schmidt independence criterion has been used as a generic measure of dependence. As a result, the proposed algorithm maximizes this notion of dependence (Chapter 5.2).

2.3 Datasets

2.3.1 EMO-DB

EMO-DB, also known as Berlin Emotional Speech dataset [21], is an acted dataset, produced by 10 professional actors (5 female and 5 male), where each has acted 10 sentences with 7 different categorical emotions: anger, boredom, disgust, fear, happiness, neutral, and sadness. The affective content of each sentence has been then judged by different referees. Although all the actors have performed the same number of sentences with the same set of emotional states, available recordings are not distributed uniformly over the 7 classes. The number of available recordings have been reduced (to 535) to maintain the quality of the dataset by not including improperly expressive utterances.

EMO-DB has been adopted in many studies including those by Amarakeerthi *et al.* [2], Burkhardt *et al.* [21], Casale *et al.* [27], Chandaka *et al.* [29], Ramakrishnan and El Emary [160], Schuller *et al.* [177–179, 183], Ślot *et al.* [191], Stuhlsatz *et al.* [198], Vankayalapati *et al.* [206], Vlasenko *et al.* [210], Wu *et al.* [226, 227], Yeh and Chi [233], and Yun and Yoo [235].

EMO-DB is the most cited affective speech dataset that comes with categorical annotation of affect, which explains the rational for adopting the dataset for analysis of categorical affect. In this thesis, this dataset is used in order to examine the notion of spectral emotion profile that is explained in Chapter 4.3.

2.3.2 VAM

VAM [92], short for 'Vera am Mittag'[1], is a spontaneous emotional speech dataset, available in both audio and video. A total number of 47 speakers, 36 female and 11 male, take part in the

[1] Vera am Mittag (Vera at noon) is a German talk show.

recordings. Their age rangers from 16 to 69 (70% of the actors are of age 35 or younger). VAM-Audio includes two modules, VAM-Audio I and II, which in total comprises about 50 minutes of recording. The division into two modules is based on the quality of emotions conveyed by speakers. VAM-Audio I, classified as very good, contains 19 speakers and in average 26.3 sentences per speaker (499 sentences in total). On the other hand, VAM-Audio II, classified as good, contains 28 speakers, where in average there is 18.5 sentences available per speaker, which adds up to 519 sentences in total. VAM-Audio is annotated using three emotional primitives: valence, activation, and dominance.

VAM dataset comes with dimensional annotation of affect based on three affective dimesions: activation, dominance, and valence, where evaluations are done based on a per-sentence basis.

Works that have adopted the VAM dataset for the study of affective speech are numerous, where some of them are conducted by Casale *et al.* [27], Espinosa *et al.* [46, 47], Eyben *et al.* [48, 49], Fewzee and Karray [53], Grimm *et al.* [75, 77], Kanluan *et al.* [92], Schuller *et al.* [173, 178, 180], Stuhlsatz *et al.* [198], Tarasov and Delany [200], Vankayalapati *et al.* [206], Vlasenko *et al.* [210], Wöllmer *et al.* [217], Siqing Wu *et al.* [226, 227], Dongrui Wu *et al.* [223, 225], Yun and Yoo [235], and Zhang *et al.* [239].

VAM dataset is the most cited dataset of affective speech recognition that comes with dimensional annotation of affect. Therefore, the multitude of the existing works on this dataset would enable us to compare the result of this study with those of the state-of-the-art in the field. In this thesis, VAM dataset is used for examining the notion of spectral emotion profile (Chapter 4.3), for dimensionality reduction (Chapter 5.1), and for assessing capabilities of the proposed max-dependence regression (Chapter 5.2).

2.3.3 SEMAINE

SEMAINE is a dataset recorded based on the sensitive artificial listener (SAL) interaction scenario [40]. The aim of SAL is to evoke strong emotional responses in a listener by controlling statements of an operator (the script is predefined in this scenario). For this purpose, four agents are introduced, where each tries to simulate a different personality: Poppy, who tries to evoke happiness, Obadiah, who tries to evoke sadness, Spike, who tries to evoke anger, and Prudence, who tries to make people sensible. A user can decide to which operator talk at any time. The combination of those decisions is therefore claimed to result in a highly emotional conversation.

SEMAINE is recorded using three different scenarios: solid-SAL, semi-automatic SAL, and automatic SAL. 150 participants (93 female and 57 male) have taken part in the recordings and their ages range from 22 to 60 (32.8 ± 11.9). Judgment of the affective contents of this dataset is done in a continuous, frame-based manner, for every 50 mSec-long window of the signal, and to assess the affective content according to four dimensions: arousal, expectancy, power, and valence.

Solid-SAL [123, 124] is a similar scenario to SAL, except there is no predefined script. Instead, operators are free to act as one of the four SAL agents at any time. This is done for the sake of a more natural face-to-face conversation. As in the SAL scenario reading the script or recalling it (in case operators have memorized the script) may not allow such non-verbal interactions.

Solid-SAL part of the SEMAINE dataset has been used as the benchmark in Audio/Visual Emotion Challenges 2011 and 2012 (AVEC'11 and '12). Therefore, despite the relative young age of the dataset, this part of the dataset has been frequently used in the literature of affective speech recognition. Some of the works that have adopted the dataset include those by Baltrušaitis *et al.* [11], Calix *et al.* [25], Cen *et al.* [28], Cruz *et al.* [34, 35], Dahmane and Meunier [37], Fewzee and Karray [54, 55], Gangeh *et al.* [63], Glodek *et al.* [68, 69], Kim *et al.* [97], van der Maaten [204], Meng and Bianchi-Berthouze [128], Nicolle *et al.* [140], Ozkan *et al.* [146], Pan *et al.* [147], Ramirez *et al.* [161], Savran *et al.* [170], Sayedelahl *et al.* [171], Schuller *et al.* [189], Soladié *et al.* [193], Sun and Moore [199], Martin Wöllmer *et al.* [220].

Given that many studies are conducted in the framework of this dataset, it has been adopted in the major part of this thesis. This choice has enabled us to compare the results of the conducted analyses and the developed algorithms with the state-of-the-art of the literature of this problem. In this thesis, the Solid-SAL part of the SEMAINE dataset is used for evaluating the proposed spectral energy distribution as a set of low-level descriptors (Chapter 4.1), for assessing the capabilities of the proposed max-dependence regression as a learning algorithm (Chapter 5.2), and for evaluating the two proposed multi-value dictionary learning algorithms (Chapter 5.3).

2.4 Conclusion

In this chapter, we reviewed the most recent studies on automatic recognition of affective speech according to their approaches in speech signal processing and statistical modeling. Furthermore, according to what we discussed in the previous chapter as to the directions of this thesis and the questions that we asked, we conclude this chapter by relating the asked questions to the recent literature of affective speech recognition.

There is a well established set of speech features that is commonly used by the majority of the studies, and there is still the tendency towards developing new features that could efficiently capture affective contents of speech. This applies to the development of new low-level descriptors, as well as new statistical functions. In addition, although different studies have included dimensionality reduction as a part of their work, in spite of the significance of variable selection, the majority of the conducted research takes the brute force approach for feature extraction. On the other hand, very few works in the literature have tried to discover the set of speech features that are most imperative to capturing affective qualities of speech, and even if they do perform dimensionality reduction of some sort, they rarely disclose their findings as to what set of features are most relevant to the recognition of affective speech.

Furthermore, in spite of the fact that individual and cultural differences are proved to impact the production and comprehension of affect, the current solutions do not take account of those differences. That is to say, according to the studies before the current research, those differences enter into the picture at the learning stage, whereas feature extraction may also closely depend on those differences. In addition, by studying the literature of regression models for predicting continuous affect, we notice that correlation coefficient is used in most, if not all, of those studies as the measure of goodness of a fit. However, the learning algorithms used in those studies, support vector machines being the most favorite one, do not maximize the correlation coefficient. Instead, a sense of prediction error is commonly used for that purpose. Finally, although sparse representation has been considered for classification, that is for the application of categorical affect, the literature of the problem does not show any track of such approaches for regression problems, i.e., modeling dimensional affect.

Chapter 3

Theoretical Background

In this chapter, we review the conventions for representing and measuring affect, and formalize the associated theoretical framework. That is, we address two subproblems that are tackled in paralinguistic speech recognition in general, affective speech recognition being a special case: 1) speech signal processing and 2) statistical modeling.

3.1 Representation of Affect

There are two different ways to describe affect: 1) to use coarse categories (e.g. anger or happiness), known as the categorical representation, 2) to use the extents of some lower level emotional attributes [94] (e.g. valence, activation, dominance), known as primitive-based or dimensional representation.

For the categorical type of emotions, a long list of emotional states are collected [33], among which only six states are known as the basic emotions. Those are anger, disgust, fear, happiness, sadness, and surprise. These are known as Ekman basic emotions, named after the psychologist Paul Ekman [42].

On the other hand, theories behind the dimensional representation claim that the space defined by those representations subsumes all the categorical emotional states. The objective of those theories is to find an efficiently sufficient emotional space. That is, the fewest number of emotional attributes or dimensions, that can describe any emotional state. The psychologist James A. Russell is one of the pioneers of the theory of affect [167]. According to Russell's Circumplex model, the cognitive structure of affect fits in a 2-D space characterized by pleasure-displeasure and arousal-sleep as the two dimensions (Figure 3.1). Another dimensional theory

Figure 3.1: Visualization of Russell's Circumplex model

is that of Mehrabian [125, 126], known as PAD. PAD stands for pleasure-arousal-dominance. According to the theory developed by Mehrabian, pleasure, arousal, and dominance are three nearly orthogonal dimensions which provide a concise representation of emotional states. A more recent dimensional view is developed by Fontaine *et al.* [57]. According to Fontaine *et al.*, to sufficiently describe emotional states in English, French, and Dutch, four dimensions are needed. Those dimensions are as follows: evaluation-pleasantness, potency-control, activation-arousal, and unpredictability.

3.2 Speech Signal Processing

The purpose of signal processing is to extract as much information as possible from the speech signal. Given that speech is a non-stationary stochastic process [168], speech samples are broken down into frames to preserve events that occur in small fractions of time. Consequently, low-level descriptors (e.g., pitch) are extracted in frame-level. That makes for a contour of values corresponding to a low-level descriptor for any given speech signal. (For such a contour, the abscissa represents time.) Since the frame-level granularity is oftentimes finer than the granularity of interest, e.g., a sentence or a few seconds, such contour is usually summarized. Descriptive statistics are commonly used for that purpose. Therefore, by doing so, we assume an underlying distribution for the contour of each low-level descriptor, and we try to capture the essence of those distributions using some sort of descriptive statistics. We refer to those statistics as statistical functional, or functionals. Therefore, each speech feature, as extracted in this manner, characterizes two (or more) pieces of information: a low-level descriptor, and a functional. In the remainder of this section, we spend some time discussing various types of low-level descriptors and functionals, as well as presenting a categorization for each of them.

3.2.1 Low-level descriptors

Speech low-level descriptors, commonly referred to as LLDs, are categorized into three classes: energy-related, spectral, and voicing-related. This categorization is based on the domain in which LLDs are extracted. Whereas voicing-related and spectral LLDs are extracted from time and frequency domains, respectively, energy-related LLDs may be extracted from either of the domains.

Energy-related

- Loudness is a psychological characterization of the physical strength of sound.
- RMS energy or Root mean squared energy is the most common definition for the energy of a signal.
- ZCR or zero-crossing rate is an indication of the energy of the signal, and is defined as the number of times that the sign of the signal changes.

Spectral

Spectral LLDs are those that are extracted from the spectrum of the speech. Various common types of spectral LLDs are as follows:

- Energy in spectral bands is defined as partial energy of the signal limited to one or more ranges of the spectrum.
- Harmonicity, also referred to as harmonic-to-noise ratio (HNR), represents the degree of acoustic priodicity, and is defined as the relative energy of the signal that can be explained by sinusoidal variations.
- MFCC, or mel-frequency cepstral coefficients, are indications of the shape of the spectral envelope of the signal, and are defined as the amplitudes of discrete cosine transform of mel-log powers of the spectrum of the signal.

- Psychoacoustic sharpness is one of the metrics of sound quality. Psychoacoustics is the principle that deals with the perception of sound.

- RASTA-filtered auditory spectrum, representing the same concept as energy in spectral bands, RASTA-filtered spectrum filters out short-term noise of the signal.

- Spectral measures. To calculate these measures, the spectrum is considered as a probability mass function.

  - Entropy. Spectral entropy is an indication of information content in the spectrum, and is defined as the Shannon's entropy of the spectrum.
  - Flux. Spectral flux is an indication of the amount of changes in the spectrum over time, and is defined as the average difference between consecutive frames' spectrum.
  - Kurotsis. Spectral kurtosis is used as an indication of the peakedness of the spectrum, as is defined as the fourth central moment of the spectrum.
  - Rolloff. Spectral rolloff is used as an indication of the bandwidth of the signal, and is defined as the frequency below which the accumulation of the signal energy passes a threshold percentage of the total energy.
  - Skewness. Spectral skewness is used as a measure of symmetry of the spectrum, and is defined as the third central moment of the spectrum.
  - Slope. Spectral slope measures the trend of the spectrum, and is defined as the linear regression coefficient of the spectrum.
  - Variance. Spectral variance is an indication of the spread of the spectrum, and is defined as the second central moment of the spectrum.

Voicing-related

- F0, commonly known as fundamental frequency, is an indication of periodicity and estimates the dominant frequency of the signal. Autocorrelation function (ACF) is commonly used to calculate F0.

- Jitter is a metric for quantifying irregularities of a quasi-periodic signal, defined as the average variation of the fundamental frequency.

- Shimmer is a metric for quantifying irregularities of a quasi-periodic signal, defined as the average variation of the energy of the signal over time.

- LogHNR is a measure of harmonicity, and is defined as autocorrelation function at the fundamental period over the difference of the autocorrelation function at zero and at the fundamental period.
- Probability of voicing measures the share of the fundamental frequency in the signal, and is defined as the quotient of autocorrelation function at the fundamental period over autocorrelation function at zero.

3.2.2 Functionals

We categorize functionals into two major classes, depending on whether they depend on the order of values in a contour or not. We refer to the first class as dynamic functionals, given that they capture the sense of dependence on time, as opposed to the second class that does not depend on time, which we refer to as static functionals.

Dynamic functionals

- First-order difference contour is used as an indicator of the rate of changes of the contour.
- Falling and rising slopes of the contour are used to describe its major trends.
- Fall and rise times are percentages of times during which the contour is falling or rising.
- Left/right-curve times indicate the major trends of the curve, and are defined as percentages of times during which the contour has left/right curvature.
- Linear predictive coefficients are used as indicators of predictability of the contour and are defined as coefficients of the finite difference equation that explains the linear dependence of the contour at each point to its past. Linear predictive coefficients are abbreviated as LPC, and are commonly referred to as linear predictive coding.
- Linear regression coefficients. are the sets of coefficients corresponding to the linear approximation of the contour, and are defined using the method of least squares.
- Quadratic regression coefficients. are the sets of coefficients corresponding to the quadratic approximation of the contour, and are defined using the method of least squares.
- Segment length is a factor that depends on the segmentation algorithm, and would be of use if static segmentation is not in place.

- Position of max and min are the relative positions of the min and max with respect to the segment length.
- Distance between peaks is an indication of the dynamicity of the signal.
- Duration quantifies the speaking rate.

Static functionals

Given that static functionals do not take into account the order of values in a contour, a contour is rather considered as a probability distribution function by these functionals. Static functionals are further categorized into two classes: those that describe the *location* of the probability distribution function, and those that describe its *dispersion* and shape.

- Location.
  - Arithmetic Mean or mean is the average value of the contour, and as defined as its first central moment.
  - Centroid is the center of gravity of the contour, as is defined as its weighted average, where sample indices are used as weights.
  - Down/up-level times are the percentage of times during which the contour is blow/above a certain percentage above the min value.
  - NNZ is defined as the percentage of the non-zero values.
  - Min and max are used to mark the extents of variations of a sample.
  - 1% and 99% percentiles are used as robust estimates of min and max of a sample, assuming that outliers fall in either of the two end-one percents of the population, and are defined as the points that separate each of the two utmost one percents of the sample from the rest.
  - Quartiles. The three quartiles, i.e., 25, 50, and 75, are the three points that equally divide a sample into four groups, each one containing one quarter of the population.
  - Root quadratic mean is a measure of the mean amplitude of a sample, and is defined as the expectation of the square of the contour.
- Dispersion.
  - Flatness indicates how noise-like the contour is, and is defined as the quotient of geometric mean over arithmetic mean.

– Inter-percentile range is used as a robust estimation of range, and is defined as the difference between 1% and 99% percentiles.

– Interquartile ranges indicate extension of the distribution in different areas by concentration of population, and are defined as differences between pairwise quartiles.

– Kurtosis indicates the degree of peakedness of a distribution and is used as a measure of non-Gaussianity. Kurtosis is defined as the fourth central moment of a distribution.

– Range of a sample is frequently used as an indication of the distribution's degree of variation, and is defined as the difference between sample's max and its min.

– Skewness is a measure of the degree of asymmetry of a distribution, and is defined as the third moment of the distribution.

– Standard deviation is a measure of the spread of the contour around its average, and is defined as the second central moment of the distribution.

3.3 Statistical Modeling

The objective of a statistical model is to find patterns of variation between extracted features (or analogously explanatory variables) from one side, and corresponding paralinguistic qualities of speech (or response variables) from the other side. In other words, at this phase, we are interested in finding a mapping f, such that

$$f : \mathbf{x} \rightarrow y, \tag{3.1}$$

where explanatory and response variables are denoted by $\mathbf{x}$ ($\in \mathbb{R}^p$) and y ($\in \mathbb{R}$), respectively. p denotes the number of explanatory variables. The function f is usually described using a set of parameters. By assuming the form of the parametric function f, the problem of statistical learning is then encapsulated by the problem of parameter estimation. In other words, we first assume that the original system using which pairs of $\mathbf{x}$ and y are generated take the same form as the function f, and then we try to estimate parameters of such model. Different algorithms, commonly referred to as learning algorithms, are used for parameter estimation, where their major difference is the criterion that each tries to optimize. A commonly practiced assumption is the linearity assumption, which dictates the following form for the function f:

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{3.2}$$

where β_0 and β_1 to β_p are parameters of the linear model, commonly referred to as linear coefficients. Estimation of the linear coefficients is commonly done by minimizing MSE or mean squared error. In this chapter, we review prevalent learning algorithms for the linear model such as ordinary least squares, elastic net, and support vector regression.

Beside the choice of model and learning algorithm, an essential part of statistical modeling is variable selection. Variable selection is performed primarily to choose a subset of explanatory variables that contribute to the goodness of a model. Additionally, by doing so, we wish to answer to the question of what are those speech features that can concisely capture affective qualities of speech. In the following, we review some dimensionality reduction algorithms such as principal component analysis, supervised principal components, supervised principal component analysis, greedy feature selection, Fished score, and Laplacian score.

3.3.1 Parameter Estimation

Ordinary Least Squares

Ordinary least squares (OLS) is among the first methods that are used for estimating parameters of linear model [81]. As the name suggests, it sets as the objective minimization of the squared error:

$$\min_{\beta,\beta_0} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2. \tag{3.3}$$

What OLS falls short of considering is taking into account different degrees of importance among explanatory variables. That is, all of them are treated as equal. This leads to difficulties such as singularity resulting from colinear explanatory variables, that in turn would result in weak analysis.

Elastic Net

Being a linear combination of the ridge regression [84] and the Lasso [201], the elastic net [202, 242] solves the following problem.

$$\min_{(\boldsymbol{\beta},\beta_0)\in\mathbb{R}^{p+1}} \left[\frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}\boldsymbol{\beta})^2 + \lambda \mathbf{P}_\alpha(\boldsymbol{\beta}) \right] \tag{3.4}$$

where

$$\begin{aligned} \mathbf{P}_\alpha(\boldsymbol{\beta}) &= (1-\alpha)\frac{1}{2}||\boldsymbol{\beta}||_{\ell_2}^2 + \alpha||\boldsymbol{\beta}||_{\ell_1} \\ &= \sum_{j=1}^{p} [\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|] \end{aligned} \tag{3.5}$$

Where n, p, and $\boldsymbol{\beta}$ are the sample size, the dimensionality of the feature space, and the parameters of the linear regression, respectively.

For $\alpha = 0$ and 1 elastic net reduces respectively to the ridge regression and the Lasso. Ridge regression is a shrinkage method and due to the smoothness of the ℓ_2 norm, it always keeps all the explanatory variables in the model. On the contrary, due to the sharp edges of the ℓ_1 constraint, the Lasso provides a compact representation of the feature space. The reason for this behavior and that of the ℓ_2 norm can been found by examining the intersection of the error contour, with respect to the regressors, and the constraints in a solution of the constrained optimization problem. In the case of the ℓ_2 constraint, the intersection can happen arbitrarily wherever on the regressors' hyperplane, whereas in the case of the ℓ_1 constraint, due to the sharp edges of the constraint at the axes, it is more likely to take place in one of the sub-spaces, which implies zero value for all the variables which are independent of the subspace.

Back to the comparison of the ridge regression and the Lasso, despite the sparse representation that the Lasso suggests, it has some limiting attributes. One is that the number of selected variables by the Lasso may not exceed the sample size and this becomes a problem when $p > n$. Also, when a group of variables are highly correlated, Lasso does not make a good selection. It is shown that in such cases, and when $n > p$, the ridge regression results in better prediction performances [201]. The elastic net is therefore proposed [242] to preserve the sparse representation of the Lasso and to alleviate its downsides.

Going back to the choice of α, as the parameter departs zero and gets closer to one, the elastic net behaves more like the Lasso than the ridge regression. For instance, by doing so, for a fixed λ, the resulting answer will become sparser. It is clear that the sparsity of the model also depends on λ: The smaller the parameter, the more number of features will be kept in the model. A proper choice of the two parameters should be set by cross validation.

From a probabilistic point of view, elastic net can be seen as a probability distribution (in the case of Equation 3.4, a Gaussian distribution) of prediction error, which is constrained with a Gaussian and Laplace prior distributions.

Support Vector Regression

Support vector regression (SVR) too is a kind of shrinkage method [14]. What makes SVR quite different than the aforementioned shrinkage method is the error that it sets to minimize; the penalty term is no different than that of the ridge regression, i.e., ℓ_2 norm.

$$\min_{\beta,\beta_0} \sum_{i=1}^{N} E_\epsilon^2 + \lambda \sum_{j=1}^{N} |\alpha_j|_{\ell_2}, \tag{3.6}$$

where α is parameter that controls which of the training samples will be used, and which will not. And, E_ϵ is an ϵ-insensitive error function, meaning that it considers errors of less than ϵ and greater than -ϵ to be zero. What this error function implies is that training samples that are within ϵ distance from the zero-error curve are disregarded [14]. Therefore, SVR takes advantage of the sparsity of training samples. Due to this reason, it is expected to be less sensitive to training set.

3.3.2 Dimensionality Reduction

Principal Component Analysis

The principal component analysis or PCA is an unsupervised dimensionality reduction method which aims at preserving the major directions of variation. To do so, it suggests a linear transformation

$$\hat{\mathbf{X}} = \mathbf{X}\mathbf{W}, \tag{3.7}$$

which satisfies the following criterion

$$\begin{aligned} \underset{\mathbf{W}}{\text{argmax}} \quad & \mathbf{W}^T\mathbf{S}\mathbf{W} \\ \text{subject to} \quad & \mathbf{W}^T\mathbf{W} = I. \end{aligned} \tag{3.8}$$

Where $\mathbf{S}$ is the covariance matrix of $\mathbf{X}$ and I is the identity matrix. Through a set of algebraic manipulations, it can be shown that for a $\hat{\mathbf{X}}$ of d dimensions, rows of $\mathbf{W}$ should be the d eigenvectors of $\mathbf{S}$, corresponding to the top d eigenvalues of $\mathbf{S}$. An alternative derivation of the PCA would be the linear transformation that minimizes the squared reconstruction error.

Supervised Principal Components

Supervised principal components or SPC, being similar to the conventional PCA, is a supervised dimensionality reduction technique. In order to make the PCA supervised, Bair and others [10] suggest a reduction of the dimensionality of the explanatory variable $\mathbf{X}$, prior to the calculation of the principal components. To do so, they select a collection of features C_θ that have regression coefficients of larger than the threshold θ.

$$C_\theta = \{i : |s_i| = \frac{|x_i^T y|}{\sqrt{x_i^T x_i}} > \theta \,; i \in \{1, \cdots, p\}\}. \tag{3.9}$$

Supervised Principal Component Analysis

Supervised principal component analysis or SPCA [13] is a dimensionality reduction method based on the Hilbert-Schmidt independence criterion (HSIC) [73]. The objective of SPCA is to maximize the dependence between a projection of the explanatory variable $\mathbf{X}$ and the response variable $\mathbf{y}$, through a projection matrix $\mathbf{W}$. To begin with an explanation of the method, let us first review the empirical estimate of HSIC.

Assuming $\mathcal{F}$ and $\mathcal{G}$ to be two separable reproducing kernel Hilbert spaces (RKHS) [83] and $\mathcal{Z} := \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\} \subseteq \mathcal{X} \times \mathcal{Y}$, HSIC($\mathcal{Z}, \mathcal{F}, \mathcal{G}$) is defined as

$$\text{HSIC}(\mathcal{Z}, \mathcal{F}, \mathcal{G}) = (N-1)^{-2}\textbf{tr}(\mathbf{KHLH}). \tag{3.10}$$

Where $\mathbf{K}$,$\mathbf{L}$,$\mathbf{H} \in \mathbb{R}^{N \times N}$, $K_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$, $L_{ij} := l(\mathbf{y}_i, \mathbf{y}_j)$, and $H := I - N^{-1}\mathbf{e}\mathbf{e}^T$ ($\mathbf{e}$ is a vector of all ones).

According to this measure and the work by Barshan and others [13], they maximize the dependence between the linear kernel of $\mathbf{XW}$ (i.e. $\mathbf{XWW}^T\mathbf{X}^T$) and a kernel of $\mathbf{y}$ by maximizing the corresponding HSIC empirical estimate.

$$\begin{aligned} \underset{\mathbf{W}}{\text{argmax}} \quad & \textbf{tr}(\mathbf{W}^T\mathbf{X}^T\mathbf{HLHXW}) \\ \text{subject to} \quad & \mathbf{W}^T\mathbf{W} = I. \end{aligned} \tag{3.11}$$

Through a set of algebraic manipulations, it can be shown that, rows of $\mathbf{W}$ should be the d eigenvectors of $\mathbf{X}^T\mathbf{HLHX}$, corresponding to the top d eigenvalues of $\mathbf{S}$.

Greedy Feature Selection

Greedy feature selection [50] is an unsupervised filter-type variable selection algorithm. At the heart of the algorithm lies minimizing the Frobenius norm of the construction error, as follows.

$$F(\mathcal{S}) = ||\mathbf{X} - \mathbf{XW}^{(\mathcal{S})}||^2_{\mathcal{F}}. \tag{3.12}$$

Where $\mathcal{S}$ denotes the set of indices corresponding to the selected subset of features, $\mathbf{W}^{(\mathcal{S})}$ does the projection of $\mathbf{X}$ onto the space spanned by the subset of selected features, $\mathcal{S}$, and $||.||_{\mathcal{F}}$ characterizes the Frobenius norm.

To minimize F, Farahat et al [50] have proposed a recursive algorithm that takes advantage of the following theorem.

$$F(\mathcal{P}) = F(\mathcal{S}) - ||\mathbf{X}(\mathbf{W}^{(\mathcal{P})} - \mathbf{W}^{(\mathcal{S})})||^2_{\mathcal{F}} \tag{3.13}$$

Where $\mathcal{S} \subset \mathcal{P}$. Then, using this theorem, they solve the following optimization problem in a recursive manner.

$$l = \underset{i}{\operatorname{argmin}} F(\mathcal{S} \cup i) \tag{3.14}$$

Where i is the index of the to-be-selected feature at each iteration.

Fisher Score

What Fisher score suggests is to select those features that make samples from the same class compact, while making samples from different classes far apart. The formulation for the Fisher score of the kth feature is as follows.

$$FS_k = \frac{\sum_{i=1}^{C} n_i(\mu_k^i - \mu_k)^2}{\sum_{i=1}^{C} n_i {\sigma_k^i}^2} \tag{3.15}$$

Where C is the number of classes, and μ_k^i and σ_k^i are the mean and standard deviation of the samples in the ith class, corresponding to the kth feature. According to the Fisher score, one may choose the set of d features that have the highest Fisher score, if d is the dimensionality of the subspace. The choice of d can be made by cross validation.

Laplacian Score

Intuitively similar to the Fisher score, the Laplacian score for the kth feature is defined as follows.

$$LS_k = \frac{\sum_{i,j}(x_{i,k} - x_{j,k})^2 s_{i,j}}{var(x_{:,k})} \tag{3.16}$$

Where $x_{i,k}$ is the values of the kth feature for the ith instance, when X is assumed as an N by p matrix containing all the data; N and p are the number of samples and the dimension of the feature space. In this equation, $var(x_{:,k})$ denotes the variance of the kth feature over all the samples. $s_{i,j}$ is an entry of an N by N matrix S, introduced as a weighting factor, accommodating the locality of the samples into the score. Given two samples $x_{i,:}$ and $x_{j,:}$ from C_i and C_j classes, $s_{i,j}$ is defined as follows.

$$s_{i,j} = \begin{cases} e^{(-\frac{||x_{i,:} - x_{j,:}||^2}{t})} & \text{if } C_i = C_j, \\ 0 & \text{if } C_i \neq C_j. \end{cases} \tag{3.17}$$

Where t is a constant that ought to be set properly.

Unlike the Fisher score, those features are more favorable here that have a lower Laplacian score.

3.4 Conclusion

In this chapter, categorical and dimensional presentations of affect are introduced, and their relationship to each other is described. Furthermore, the process of extracting features from speech signal is briefly reviewed, and it is clarified how speech features consist of a low-level descriptor and a functional. This is followed by a list of such low-level descriptors and functionals that are most commonly used in the literature of affective speech recognition, which are as well utilised in this thesis. Additionally, reviewed in the chapter are the prevalently used machine learning algorithms that have been either used in the literature of the problem, or have been exploited throughout this thesis. Finally, the three most cited affective speech datasets are introduced. These datasets have been used in different experiments in this thesis.

Chapter 4

Proposed Speech Signal Processing Solutions

4.1 Sepectral Energy Distribution

Feature extraction and dimensionality reduction comprise the most imperative parts of emotional speech recognition problem. As a solution to this problem, we propose a new set of speech features based on the distribution of energy in frequency domain. The proposed set of features are distinguished from previous works particularly by the degree of freedom that they offer. This degree of freedom is shown to be advantageous for extracting features tailored to the learning problem. To investigate the applicability of the proposed set of speech features, we have performed experiments in the frameworks of international audio/visual emotion challenge (AVEC) 2011, 2012, and Interspeech 2012.

4.1.1 Background

A major motivation for this work has been shaped towards achieving the objective of the international audio/visual emotion challenges and workshops (AVEC) [188,189]. In the context of three sub-challenges, participants were expected to predict emotional contents of speech in word level and fully continuous granularities and based on four binary dimensions: activity, expectation, power, and valence. This enables us to make a comparison between our approach and the most recent advances in the field. In the following, we go briefly over some of the works done by the participants in the audio sub-challenge.

In order to deal with feature extraction and dimensionality reduction, Calix et al [25] have used a Chi-square ranking process to select a subset of features. In a work by Cen et al [28], one can see the use of the ℓ_1 norm, employed in a regression formulation, to select an optimal subset of speech features. Glodek et al [69] employed a multiple classifier system for recognizing emotional states, based on voice and facial cues. Focusing mainly on dimensionality reduction, Kim et al [97] apply maximum average recall, maximum relevance, and minimal-redundancy-maximal-relevance for feature selection. Pan and others [147] have used PCA to reduce dimensionality, along with SVM and AdaBoost for learning. Proposed by Sayedelahl et al [171] is two sets of speech features: One based on the co-occurrence matrix for the extraction of meta features from speech, and the other based on a sense of energy distribution in the frequency domain. In a work by Sun and Moore [199], the applicability of two new sets of features, namely glottal waveform parameters and Teager's energy operators, has been investigated.

According to the literature of the problem, the major focus is directed towards the statistical modeling aspects of the problem, leaving feature extraction at a secondary level of importance. However, the success of a statistical model relies significantly on the optimality of the selected set of speech features. When we say optimal, we mean to address two characteristics of a set of features: to be least in number, but most in conveyance. In this study, we pursue this objective in two ways: (1) by introducing a new set of speech features, which we name spectral energy distribution, or briefly SED. And (2) by reducing the dimensionality of the feature vector using the lasso. To verify the applicability of the proposed set of features to affective speech recognition, we have done an experiment in the framework of the first international audio/visual emotion challenges (AVEC) 2011 [189] and 2012 [188], and Interspeech speaker trait challenge (ISTC) 2012 [187]. This selection of datasets allows us to examine SED for recognition of categorical affect (AVEC'11), dimensional affect (AVEC'12), and recognition of speaker trait (ISTC'12).

4.1.2 Spectral Energy Distribution

Spectral energy distribution (SED) [172], as we define, is composed of a set of components. For a continuous-time speech signal s_t, we define the component i as follows.

$$\text{SED}_t^i = \int_{l_i}^{u_i} |g(S_\omega)|^2 d\omega. \tag{4.1}$$

Where S_ω is the Fourier transform of the speech signal s_t. And $g(.)$ is a *normalizing* function, on which we will later shed light (4.1.3). For now, let's assume $g(S_\omega) = S_\omega$.

Therefore, SED^i is the energy of the speech signal confined within the spectral interval $[l_i, u_i]$. We refer to the lower and upper bounds of a spectral interval as the interval's (or component's) parameters. Therefore, when put all the components of SED together, we will have a binned power spectrum of the speech signal.

The discrete-time formulation for the SED, given a speech signal $s[n]$, is as follows.

$$\text{SED}_t^i = \frac{1}{N}\sum_{k=1}^{N}(H[k-L_i] - H[k-U_i])|g(S[k])|^2. \tag{4.2}$$

Where $S[k]$ is the discrete Fourier transform (DFT) of the speech signal $s[n]$ and $H[k]$ is the step function. L_i and U_i are analogous to the l_i and u_i, as defined for the continuous case, and they denote the lower and upper bounds of the component i.

Figure 4.1 demonstrates extracted SED components from a speech signal, for an arbitrary choice of component parameters.

As follows, we talk about the normalizing function $g(.)$.

4.1.3 Normalizing Function

By carefully looking at Figure 4.1 we observe that some areas in the frequency domain that have more share in the total energy of the signal than some others. This in turn will hold for SED components. That is, there are some components which take greater values than some others. However, this does not mean that components with greater values are of more importance in capturing the emotional contents of speech, than those with smaller values. Therefore, it is required to normalize the components.

To normalize the components, we make use of two characteristics, one of the speech signal and one of the discrete Fourier transform. About the magnitude of a speech signal, we know that $0 < |s[n]| < 1$. And given this fact, and according to the DFT, one can prove that $0 < |S[k]| < 1$. This will hence (for $0 < p < 1$) guarantee the validity of the following inequality.

$$0 < |S[k]| < |S^p[k]| < 1. \tag{4.3}$$

As $p \to 0$, $S^p[k] \to 1$, therefore by setting the exponent to a small value, the range of variation for $S[k]$ will shrink. Accordingly, by setting

$$g(S[k]) = S^p[k] \tag{4.4}$$

in 4.2, the range of variation of the resulting SED components will become smaller. Figure 4.2 shows the effect of this normalization on SED components, for different values of p.

As follows, we discuss computational complexity of extracting SED.

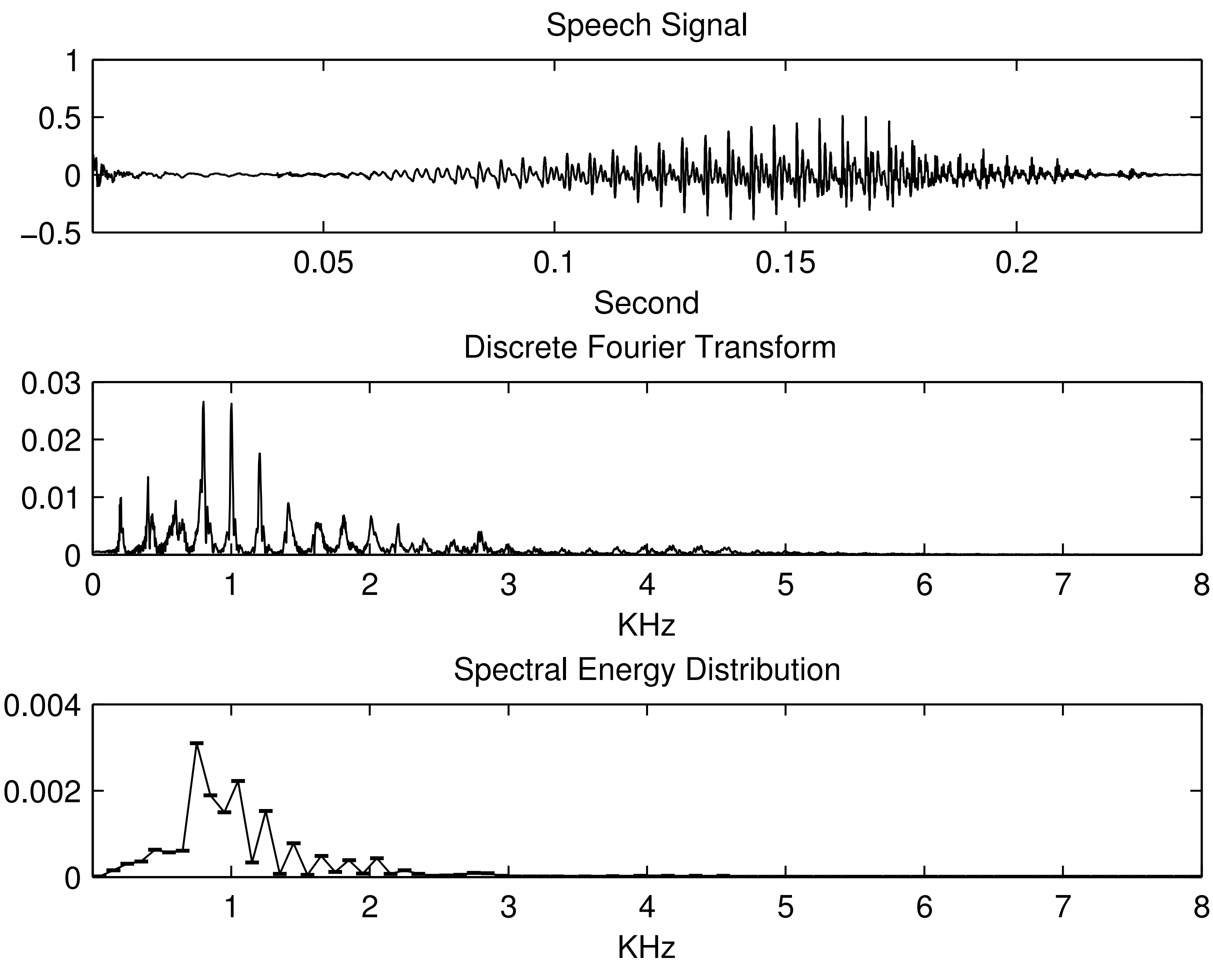


Figure 4.1: A piece of speech signal, its DFT, and the extracted SED components

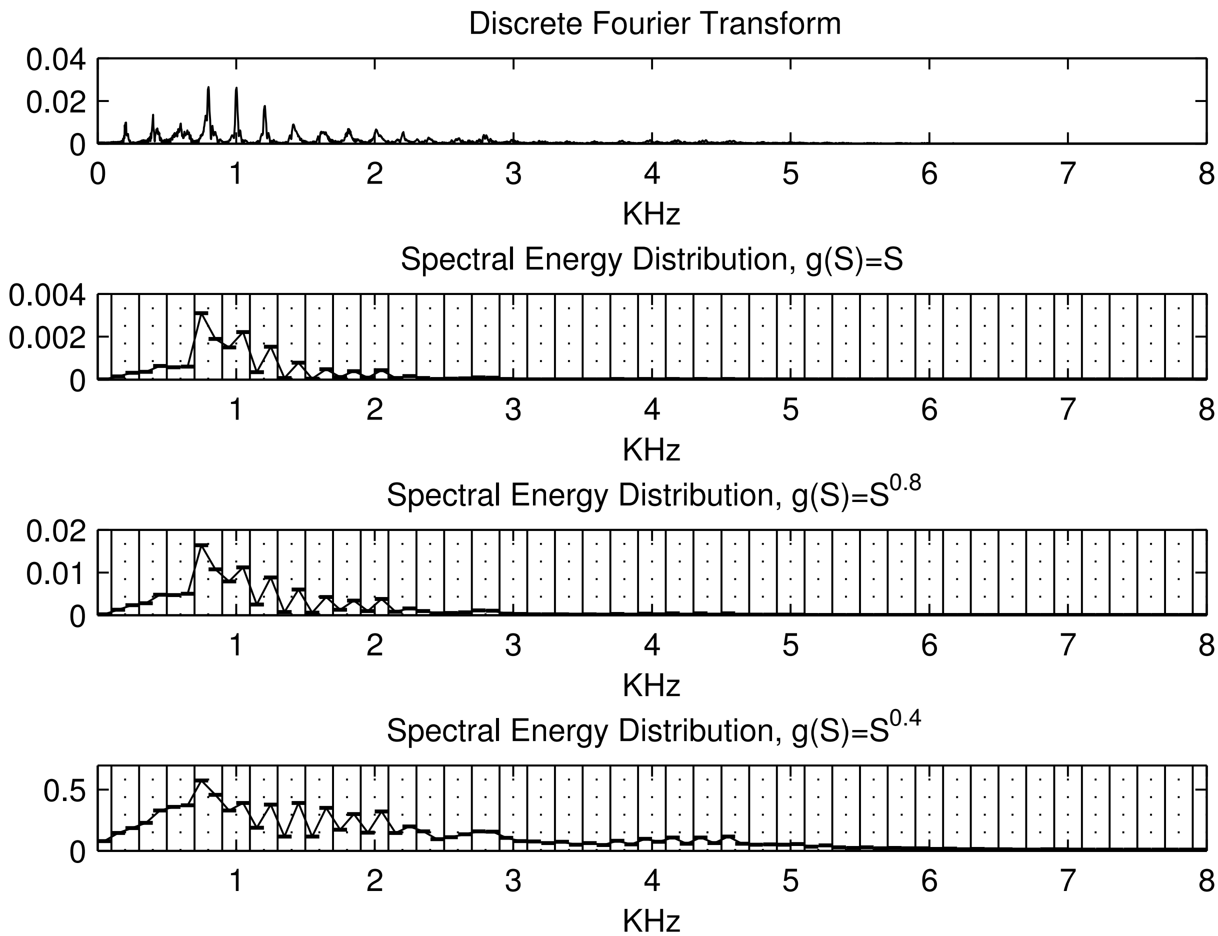


Figure 4.2: Normalization using rational exponent

Table 4.1: A comparison of required number of floating point operations for calculating SED and MFCC (zero-order terms are neglected) (N: length of speech signal in samples)

	Analytic		Numerical	
	addition	multiplication	addition	multiplication
SED	$9N^2 + 3N$	$14N^2 + 2N$	5,762,400	8,961,600
MFCC	$18N^2 + 20N$	$28N^2 + 20N$	11,536,000	17936000

4.1.4 Computational Complexity

We consider the computational complexity of the extraction of SED compared to mel-frequency cepstrum coefficients (MFCC), as one of the most commonly used LLDs. The asymptotic discussion of the computational complexity (the big $\mathcal{O}$ notation) may not fit the nature of these low-level descriptors, as the extraction is commonly done based on the segments of a speech signal. Therefore, the size of the problem may not exceed a fairly small fixed number N, which will in turn leave no place for a discussion of asymptote. N is the size of the speech signal, which is limited to the size of the segments and the sampling frequency. Therefore, N is equal to the segment time multiplied by the sampling frequency. To give a sense of the number, the segment size usually does not top 50 mSec, and a common sampling frequency is 16 kHz, which in turn results in $N = 800$. Therefore, we perform comparisons based on the number of floating point operations that takes to extract each descriptor for a fixed-size problem. The major difference between the two approaches is the neglect of the lower order terms as $N \rightarrow \infty$, which can make a difference for a small size problem.

Table 4.1 shows the number of floating point additions and multiplications required for extracting MFCC and SED. In this table, $N = 800$ is considered for the numerical comparison. According to this comparison, we can see that extraction of MFCC takes more than twice as many as the number of operations that it takes to extract SED.

4.1.5 Experimental Study: Binary Dimensional Affect

Setup

In this experiment, the solid-SAL part of the SEMAINE dataset [123] is used. Estimating affective contents of linguistic words are of interest in this experiment. The four affective dimensions considered in this experiment are as follows: activity, expectancy, power, and valence. Therefore, the objective of this experiment is to classify spoken words to extreme extents of each of the four dimensions.

Table 4.2: A comparison among some of the most recent works in the field, all performed in the framework of AVEC 2011. WA stands for weighted average (average of recall accuracy in percentage) and NF for number of features used for the task.

	Activation		Expectancy		Power		Valence		
	WA	NF	WA	NF	WA	NF	WA	NF	Average
Kim et al [97]	65.1	N/A	54.3	N/A	61.3	N/A	64.0	N/A	61.1
Calix et al [25]	63.8	1273	63.5	363	65.0	652	57.0	714	62.3
Schuller et al [189]	63.7	1941	63.2	1941	65.6	1941	58.1	1941	62.7
Cen et al [28]	58.7	≈1000	66.5	≈1000	65.9	≈1000	62.9	≈1000	63.5
Pan et al [147]	**65.4**	1941	66.5	621	64.5	1941	63.5	1941	65.0
Sayedelahl et al [171]	61.1	125	66.6	85	67.4	125	65.4	125	65.1
This work	61.3	**1**	**66.7**	**5**	**67.4**	**5**	**66.1**	**6**	**65.4**

As for statistical model and learning algorithm, we use a linear classifier, and we estimate its coefficients using lasso, assuming that the underlying distribution of the data is binomial. Furthermore, we use the partitioning of the data as was used in the audio/video emotion challenge 2011 [189]. That is, we use the training portion of the data for training purposes, including cross validation for setting the hyper-parameter of lasso. We then use the development portion of the data for hypothesis testing purposes.

As for the extraction of SED, having performed a linear search, we have set the length of the spectral intervals equally to 100 Hz. Spectral intervals do not intersect and they cover 0 to 8 kHz (i.e. 80 intervals). The power of $g(.)$, as in Equation 4.4, is set to 0.2. Extraction is done in a local fashion, meaning that each of the features has been extracted from 100 mSec-long windows of signal. Statistics of the extracted features over the windows is then calculated. As for the choice of statistics, we have chosen minimum, maximum, mean, median, and standard deviation. The resulting feature vector is of a length 400.

Results and Discussion

The result of this experiment is shown on Table 4.2. For comparison purposes among participants in AVEC 2011, we have picked those papers which report their weighted average (WA) on the development set, rather than the unweighted average (UA). According to Table 4.2 , the proposed model for emotional speech shows the best overall result in comparison to all the chosen participants, from both accuracy and complexity perspectives. From accuracy point of view, we can see the better performance of the proposed model for expectancy, power, and valence dimensions, as well as the mean WA over the four dimensions.

Table 4.3: The proposed set of features for modeling emotional contents of speech for each of the emotional dimensions, along with the corresponding spectral intervals.

Emotional Dimension	Activation	Expectancy	Power	Valence
Employed Features	mean SED^{29}	max/mean SED^{1} max/mean SED^{14} max SED^{40}	mean/med SED^{1} max/mean SED^{28} mean SED^{30}	mean SED^{1} med SED^{30} max/med SED^{32} mean/med SED^{47}
Corresponding Spectral Intervals	2.8–2.9 kHz	0–0.1 kHz 1.3–1.4 kHz 3.9–4 kHz	0–0.1 kHz 2.7–2.8 kHz 2.9–3 kHz	0–0.1 kHz 2.9–3 kHz 3.1–3.2 kHz 4.6–4.7 kHz

From complexity point of view, we can see how minimal the proposed model is. On the one hand, the proposed set of features includes not more than 15 features, whereas the length of feature vectors of other works ranges from 210 to more than 2500. On the other hand, this set of 15 features is composed of an explicit set of features described by SED. The proposed set of features come associates with 8 different spectral intervals of size 100 Hz. Table 4.3 gives a listing of the features and the intervals. According to this table, the use of the first spectral interval (SED^1) has seen for modeling expectancy, power, and valence dimensions. One may also notice that the intervals around 3 kHz have shown good capability in preserving emotional content of speech.

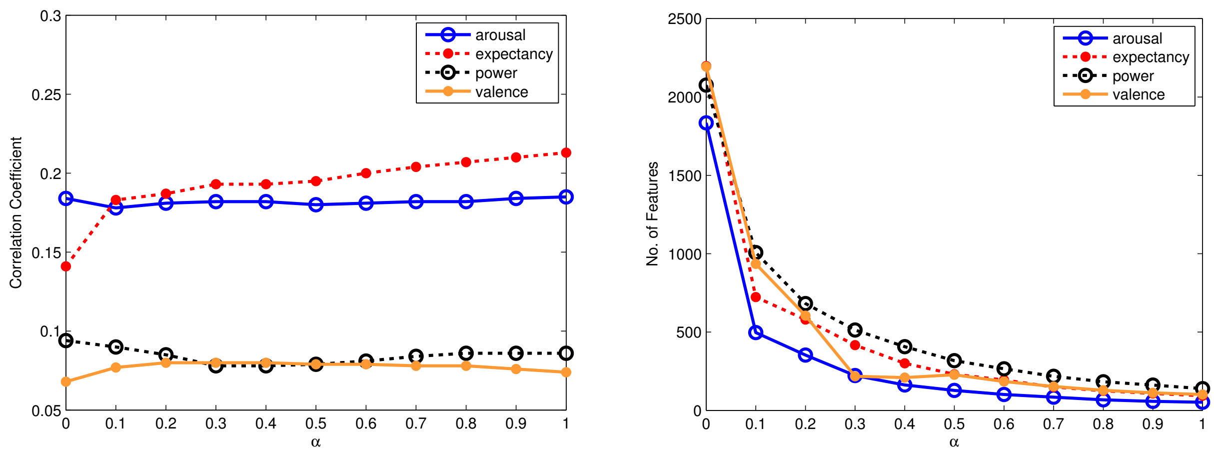
It is worthy to notice that Cen and others [28] have also used the lasso as for the choice of feature selection. However, they have run their study based on the baseline features of the challenge [189]. The baseline feature vector is a vector of a size 1941, composed of a variety features, including energy and voicing related, as well as spectral features. This can be due to two reasons. One is the sub-optimality of selection algorithms and that it becomes more severe as the size of a feature vector grows. Also, the informativeness of individual features in two feature vectors can contribute to this happening.

4.1.6 Experimental Study: Continuous Dimensional Affect

Setup

Solid-SAL part of the SEMAINE dataset is used in this experiment. The two granularities of interest are fully continuous and word-level. These correspond to the two sub-challenges addressed in the second audio/visual emotion challenge (AVEC 2012) [188]. The objective of the fully-continuous experiment is to predict affective content of speech in a continuous manner, 50-mSec long frames, whereas that of the word-level is to predict affective contents of spoken words. Four affective dimensions considered in this experiment are as follows: arousal, expectation, power, and valence.

As for statistical model and learning algorithm, we use a linear model, and estimate its coefficients using elastic net, assuming that the underlying distribution of the error is Gaussian. Furthermore, we use the partitioning of the data as was used in the audio/video emotion challenge 2011 [188]. That is, we use the training portion of the data for training purposes, including cross validation for setting the hyper-parameters of elastic net. We then use the development and test portions for hypothesis testing purposes.



(a) (b)

Figure 4.3: Correlation coefficients and the number of features against the elastic net parameter α, for the case where `a+b` is employed for the FCSC. (This figure is provided as an example, to show the trend of changes of the correlation coefficient and the number of features with the changes of α.)

Three sets of features are used in this experiment:

`a` . This set of features is composed of SED components. Extraction is done from 100 mSec windows of speech signal and the length of the spectral intervals is set to 100 Hz and they cover 0 to 8 kHz. The exponent p, as in Equation 4.4, is set to 0.2. To set each of these parameters, a line search is performed. As for the statistics, we use the min, max, median, mean, and standard deviation of the features over windows of a speech signal. This feature vector is of 400 dimensions. We use this set of features for both WLSC and FCSC.

Table 4.4: Prediction accuracy on the development set (A: audio, V: video; CC: correlation coefficient, NF: number of features; EN: elastic net, SVM: support vector machine)

Feature Set	Learning Method	Arousal CC	Arousal NF	Expectation CC	Expectation NF	Power CC	Power NF	Valence CC	Valence NF	**mean** CC
Fully Continuous Sub-Challenge										
a	EN	0.115	15	0.065	25	0.029	214	0.083	19	0.073
b	EN	0.181	51	0.213	95	0.094	1726	0.079	409	0.142
a+b	EN	0.185	52	0.213	94	0.094	2076	0.080	210	0.143
Word-Level Sub-Challenge										
a	EN	0.066	21	0.029	102	0.006	159	0.060	16	0.037
b	EN	0.066	132	0.065	114	0.004	228	0.074	213	0.052
	SVM [188]	0.097	1841	0.052	1841	0.061	1841	0.085	1841	0.074
a+b	EN	0.071	58	0.073	141	0.007	202	0.077	183	0.057

Table 4.5: Elastic net parameters (as in Equations 3.4 and 3.5), corresponding to the results shown in Table 4.4

Feature Set	Arousal α	Arousal λ	Expectation α	Expectation λ	Power α	Power λ	Valence α	Valence λ
Fully Continuous Sub-Challenge								
a	1	0.0086	1	0.1163	0.1	0.0024	1	0.0043
b	1	0.0103	1	0.2228	0	0.0341	0.3	0.0072
a+b	1	0.0102	1	0.2271	0	0.0374	0.4	0.0102
Word-Level Sub-Challenge								
a	1	0.0067	0.2	0.0987	0.1	0.0028	1	0.0030
b	0.4	0.0084	1	0.1012	0.7	0.0021	0.1	0.0296
a+b	0.8	0.0075	0.9	0.1016	0.9	0.0019	1	0.0030

- `b` . This set is composed of the provided feature set as the baseline feature vector and it is of 1841 dimensions [188]. We use this set of features for both WLSC and FCSC.

For the FCSC, we have used the elastic net criterion to train three linear models for each of the four emotion primitives, each time using one of the `a`, `b`, and `a+b` feature sets. On the other hand, for the WLSC, we have used the elastic net criterion to train four linear models for each of the four emotion primitives, each time using one of the `a`, `b`, and `a+b` feature sets. To do the training, we comply with the AVEC 2012 conditions. That is, to use the provided training set for training purposes and predict the emotional contents of the provided development set. Therefore, the development set is not seen in the training stage. The measure of comparison for this experiment is the Pearson's correlation coefficient (CC) of predicted and actual values.

Results and Discussion

Results of this experiments is presented in Table 4.4. In this table, we have also included the baseline results. Table 4.5 lists the parameters of the elastic net, α and λ, for each one of the tasks, which are set in the process of learning. Figure 4.3 shows how the prediction accuracy and the sparsity of the feature vector change with the changes of α.

The results of both FCSC and WLSC experiments, although offer a relatively sparse representation of the feature sets, do not show improvement over the baseline. Except the arousal and expectation dimensions for the FCSC, for which the elastic net results in a more accurate prediction compared to SVM, for the other regression tasks, the contrary is the case. Particularly for those two regression tasks (FCSC arousal and expectation), the size of the feature vector resulting from the elastic net is considerably smaller (less than 1%) than the feature vectors used as the baseline. According to the results of this study, SED-based features show to complement the baseline features of the challenge, considering that the `a+b` set of features outperforms both `a` and `b` sets in most of the studied cases.

4.1.7 Experimental Study: Speaker Trait

Table 4.6: Prediction accuracy on the development set [187] (UA: unweighted average, NF: no. of features; EN: elastic net, SVM: support vector machine, RF: random forests, 6K: 6125; O: openness, C: conscientious, E: extraversion, A: agreeableness, N: neuroticism)

Feature Set	Learning Method	Personality										Likability		Intelligibility	
		O		C		E		A		N					
		UA	NF	UA	NF	UA	NF	UA	NF	UA	NF	UA	NF	UA	NF
a	EN	58.3	**60**	68.5	**31**	78.1	**34**	**70.3**	**17**	70.7	**7**	**67.7**	32	60.9	322
b	EN	**62.7**	89	74.3	98	85.8	54	64.5	144	74.3	203	61.8	**18**	62.3	**227**
	SVM [187]	60.4	6K	74.5	6K	80.9	6K	67.6	6K	68.0	6K	58.5	6K	61.4	6K
	RF [187]	57.7	6K	**74.9**	6K	82.8	6K	67.2	6K	68.9	6K	57.6	6K	**65.1**	6K
a+b	EN	61.1	85	74.3	98	**86.4**	318	69.2	23	**74.7**	206	61.9	212	61.8	353

Setup

This experiment is conducted in the framework of Interspeech 2012 speaker trait challenge [187]. The problem that we deal with in this experiment is to classify extreme extents of seven trait dimensions, including 5 personality dimensions, as well as likability and intelligibility measures of human trait. In this experiment, three datasets used for the personality, likability, and intelligibility challenges are SPC, SLD, and NCSC respectively [187].

Table 4.7: Elastic net parameters (as in Equations 3.4 and 3.5), corresponding to the results provided in Table 4.6 (O: openness, C: conscientious, E: extraversion, A: agreeableness, N: neuroticism)

Feature Set	Personality										Likability		Intelligibility	
	O		C		E		A		N					
	α	λ	α	λ	α	λ	α	λ	α	λ	α	λ	α	λ
`a`	1	0.0065	0.6	0.0389	0.4	0.0548	0.9	0.0236	1	0.0487	0.8	0.0280	1	0.0007
`b`	1	0.0401	0.3	0.1660	0.6	0.1058	0.2	0.2252	0.1	0.4747	1	0.1289	0.2	0.2069
`a+b`	1	0.0414	0.3	0.1660	0.1	0.3263	0.1	0.0731	0.2	0.1808	1	0.0293	0.1	0.3249

Table 4.8: Prediction accuracy (UA) of the test set [187] (O: openness, C: conscientious, E: extraversion, A: agreeableness, N: neuroticism)

	Personality					Likability	Intelligibility
	O	C	E	A	N		
baseline [187]	58.8	80.1	75.3	**64.2**	**64.5**	59.0	69.6
this work	**60.1**	**81.6**	**78.4**	63.9	56.1	**59.3**	**70.0**

As for statistical model and learning algorithm, we use a linear model, and estimate its coefficients using elastic net, assuming that the underlying distribution of the error is Gaussian. Furthermore, we use the partitioning of the data as was used in the Interspeech 2012 speaker trait challenge [187]. That is, we use the training portion of the data for training purposes, including cross validation for setting the hyper-parameters of elastic net. We then use the development and test portions for hypothesis testing purposes.

Two sets of features are used for this experiment.

`a` . This set is composed of SED components, extracted from 100 mSec long windows of signal, the length of spectral intervals are set to 100 Hz and they cover 0-8kHz, and p as in Equation 4.4 is set to 0.2. All the parameters are optimized by performing line search. Using min, max, median, mean, and standard deviation as for the statistics of the features over windows, makes this feature vector of 400 dimensions.

`b` . This set is composed of the challenge baseline features and is of 6125 dimensions [187].

Using the elastic net criterion, three linear models are trained for each of the seven dimensions, at each time using one of the `a`, `b`, and `a+b` feature sets. As for the conditions of the experiments we abide by the Interspeech 2012 speaker trait challenge [187] rules. That is, provided training set of speech samples is used to train the model and the evaluation of the performance of the prediction is done according to the development set. Therefore, no use of the development set is done at the training phase. The measure of comparison for this experiment is the unweighted average of prediction accuracy, shown as UA.

Results and Discussion

The resulting prediction accuracies on the development set are shown in Table 4.6. In this table, we also see the baseline results of the challenge on the feature set `b`, performed by support vector machines (SVM) and random forests (RF). Table 4.7 lists the parameters as set for the training of those models according to Equations 3.4 and 3.5. Table 4.8 provides the prediction accuracy of the experiment on the test set, put together with the best (between SVM and RF) prediction accuracy of the baseline paper on the same set.

According to the Table 4.6 we notice that the number of features that is employed by training using elastic net is far less than the number of features that are used for learning with either support vector machines or random forests. For instance, the number of features that is used from the feature set `b` for modeling the five personality dimensions ranges from 54 to 203, with the average of 118. This number is less than two percent of the number of features that is used from the same feature set for modeling using SVM and RF. This is when the prediction performance of the elastic net is even higher by two percents than those of the SVM and RF. On the other hand, the result of prediction using the feature set `a` shows comparable accuracy to those performed by the feature set `b`, and still with considerably smaller size of feature vector. The overall accuracy of the prediction for the personality dimensions shows that in average the best performance is obtained by the elastic net, when the feature set `a+b` is used. As another example, the best prediction accuracy for the likability dimension is obtained by the elastic net on the feature set `a` and yet the number of features employed for this task is significantly less than the other four cases. The number of features used for this task is 32 (almost half a percent of the size of the feature set `b`), however the obtained prediction accuracy is about ten percent higher than those performed on the feature set `b`. The best prediction accuracy obtained for the intelligibility is that of the RF on the feature set `b`.

Among the five personality dimensions (Table 4.6), we can see that for openness and agreeableness the prediction accuracies obtained by one of the two `a` and `b` is higher than those of their combinations. This suggests that the elastic net has been trapped in a local optimal solution. This is as well seen for the likability dimension. Based on this observation we can conclude that, in spite of dimensionality reduction, the combination of two or more feature sets does not necessarily result in a more accurate prediction compared to a case where just one of the sets is employed. This is mainly due to the sub-optimality of the dimensionality reduction algorithms. Therefore, a longer vector of features may not necessarily give out a more descriptive model.

4.2 Measure of Dynamicity

Considering the dynamic nature of speech, a sufficient set of features should take account of those characteristics that maintain that nature. Proposed in this section is a speech feature that measures rate of changes of speech. The suggested measure, that is based on the definition of the spectral energy distribution and the KL divergence, quantifies the statistical differences of the spectrum of the signal over periods of time, as a measure of dynamicity. As for the case studies, we have adopted the Interspeech 2013 Computational Paralinguistics Challenge.

4.2.1 Background

Regardless of the type of an LLD, extraction is usually done from short windows of the signal. Therefore, to summarize the contour of the extracted features for each speech sample, some sorts of statistics of those contours are used. Although the use of statistical measures for describing such contours is inevitable, that comes with a price and that is the loss of some information that may not be captured by those measures. The purpose of this study is to suggest a new measure that can decrease the amount of lost information caused by the use of statistical measures for summarizing the distribution of features. The proposed measure takes advantage of spectral energy distribution as a set of low-level descriptors, as proposed in the previous section. To take into consideration the dynamic nature of the speech signal, the proposed measure takes account of the changes of speech in time, by estimating the relative changes of the spectrum of the signal.

To investigate the applicability of the proposed speech feature, we have adopted the Interspeech 2013 Computational Paralinguistics Challenge (ComParE). This challenge targets a set of different recognition problems regarding the paralinguistics of speech. Those problems are conflict, emotion, and autism. The conflict sub-challenge is aimed at identifying conflict in group discussion; for the emotion sub-challenge, the objective is to recognize high/low arousal, and positive/negative valence; and the autism sub-challenge is aimed at recognizing autism from speech. Aligned with the ComParE, to examine the informativeness of the proposed feature, we discuss the relevance of the proposed measure to the recognition of conflict, emotion, and autism.

4.2.2 Measure of Dynamicity

To capture the dynamic nature of the speech signal, features are commonly extracted from short windows of the signal. For the speech signal $s[n]$, let us denote a feature extracted from the window j by z^j. Therefore, to summarize the feature over a number of windows W, some descriptive statistics of the corresponding values are used. As for the choice of statistics, those that describe the center and the spread of the distribution are commonly used. Although those types of measures are essential for encapsulating a distribution, they can not sufficiently reflect the temporal nature of the signal. That is to say, changes of signal from one window to another, and changes along the time axis in general, may not be captured by them. Therefore, in addition, other types of statistical measures such as linear prediction coefficients are used. Considering this and based on the definition of spectral energy distribution (SED) [54, 55], we propose a new measure for dynamicity.

Spectral energy distribution specifies a probability mass function for the energy of speech over short windows of the signal. On the other hand, considering the dynamic nature of speech, changes of the spectral energy distribution from one window to another can be a good measure for quantifying the dynamicity of speech. Therefore, for this purpose, that is to quantify the rate of changes of speech, we use the KL divergence of SED as follows. For a speech signal with W windows, for any two consecutive windows, where $0 < j < W$, we define dynamicity as

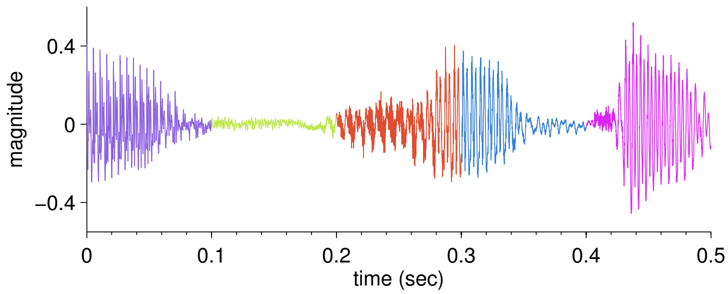
$$D_{KL}(\text{SED}^j||\text{SED}^{j+1}). \tag{4.5}$$

Where D_{KL} is the KL digergence. KL divergence, a short hand for Kullback-Leibler divergence, is a measure of difference between two probability distribution functions [106]. For two probability mass functions P and Q, KL divergence is defined as follows.

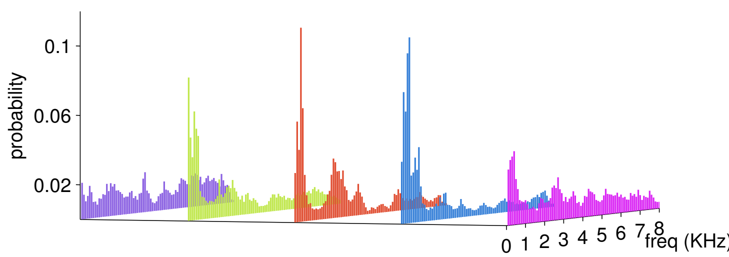
$$D_{KL}(P||Q) = \sum_i P(i)\ln(\frac{P(i)}{Q(i)}) \tag{4.6}$$

Based on this definition, the more similar two distributions are, the less their KL divergence will be. In other words, the KL divergence of two identical distributions is zero, and that is the minimum value that D_{KL} may take. Alternatives to KLD, for measuring the distance between two probability distribution functions, are Jensen-Shannon divergence, Kolmogorov-Smirnov test, Chi-Squared test, Cramer-von Mises criterion, Anderson-Darling test, and histogram intersection kernel.

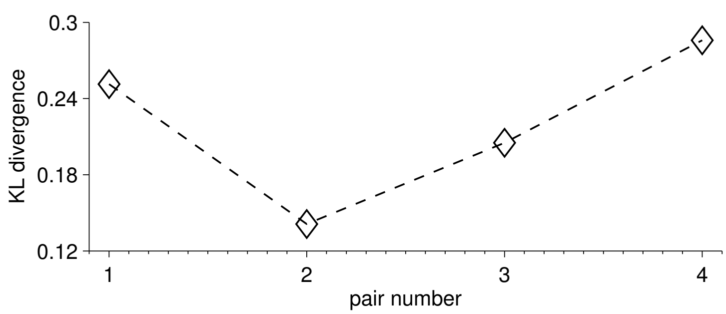
Later on, to describe the dynamicity of a piece of speech, we can use some descriptive statistics of this distribution. Figure 4.4 shows a speech sample with five windows, along with the spectral energy distribution corresponding to each window, and the KL divergence of the consecutive windows' SED.



(a)



(b)



(c)

Figure 4.4: (a) A sample speech signal. Various colors are used to show different windows of the signal. (b) Spectral energy distribution corresponding to the different windows of the signal. (c) KL divergence of the spectral energy distribution of consecutive windows of the signal, as a measure of dynamicity.

4.2.3 Experimental Study

Setup

In this experiment, we use a set of paralinguistic recognition tasks to evaluate the efficacy of the proposed measure of dynamicity. Those include recognition of conflict, affect, and autism from speech. These experiments are performed in the framework of Interspeech 2013 Computational Paralinguistics Challenge [176]. The objective of this experiment is to classify phrases according to different criterion: whether there is a conflict or not, whether the speaker is an autistic patient, whether the speaker is highly aroused, and whether a spoken phrase of high or low valence.

Our choice of model family and learning algorithm are linear models and lasso, respectively. Although the linearity assumption might be a naïve one, we use linear models due to the simplicity of those models, and the fact that they are more resistant to over-fitting. Moreover, given that the baseline experiments are conducted based on linear models, it enables us to perform a fair comparison between the utilized features in this study and the baseline features of the challenge. Furthermore, we use the train and development partitioning of the data, according to the framework of the challenge. Therefore, to fix model parameters, as well as hyper parameters, we use a 10-fold cross validation on the train set, and we use the development set for comparing purposes.

As for the low-level descriptors, three different sets of features are used. Those are as follows:

- `a` This set is composed of the components of the spectral energy distribution, extracted from 100 mSec windows of speech. We have set the length of each component to 100 Hz and together they cover 0-8 KHz. This range is limited to the sampling frequency, i.e., the Nyquist theorem. The components are chosen to be non-overlapping, therefore there are 80 of them in total. As for the statistics, we use the mean and the standard deviation of the distribution of each component, as well as the quartiles of each of them. The dimensionality of this feature set is 560.

- `b` This set is composed of the statistics of the dynamicity measure contour. As for the statistics, similar to the set `a`, we use the mean, standard deviation, and the quartiles of the distribution. The dimensionality of this set is 7.

- `c` This set of features comprises the baseline features of the challenge [176]. The dimensionality of this set is 6373.

Here, we compare the prediction accuracy and complexity of different models for each of the three sub-challenges, to those of the baseline paper on the challenge. The comparison of

Table 4.9: Prediction results on the development set. (UAR: unweighted average recall, NF: number of features)

	Set `b`		Set `a`		Set `a+b`		Set `c`	
	UAR	NF	UAR	NF	UAR	NF	UAR	NF
Conflict Sub-Challenge								
Class	71.5	7	77.5	54	**79.9**	**55**	76.4	130
Emotion Sub-Challenge								
Arousal	75.6	6	80.6	39	79.3	28	**82.7**	**54**
Valence	55.8	3	65.9	59	65.9	62	**69.7**	121
Autism Sub-Challenge								
Typicality	60.7	7	85.4	144	83.2	153	**90.4**	245

accuracy is done by means of the unweighted average recall of the prediction, whereas as for the complexity measure, we use the number of regressors or features used for each model. As for the variety of models in this comparison, we use four different sets of features: set `b`, `a`, the combination of the two sets, that is `a+b`, and the set `c`. This choice of various models make a comparison among the different sets of features more realistic, as the choice of statistical model and learning algorithm is preserved in all the cases. Table 4.9 includes the numerical results of this comparison.

Results and Discussion

Regarding the feature set `b`, other than the conflict and arousal dimensions, for which it makes possible a relatively good separation, for the other dimensions, the use of that set all by it self may not be sufficient. For the conflict sub-challenge, the feature set `a+b` proves to result in the best prediction accuracy among all competitors, including the baseline work. For the arousal dimension of the emotion sub-challenge, the feature set `c` results in the highest accuracy. For the rest of the tasks, which are the valence and categories of the emotion sub-challenge, and typicality and diagnosis of the of the autism sub-challenge, the accuracy of our results do not meet those of the baseline results. An important point to note here is that for emotion categories and autism diagnosis, feature set `a+b` results in a better predictions than those of the feature set `c`.

Figures 4.5, 4.6, and 4.7 are to suggest an interpretation of the measure of dynamicity in the context of three different paralinguistic problems. For plotting these, among the seven features of the set `b`, we have used the most discriminating features. Based on the Figure 4.5, high conflict recordings are characterized by higher minimum and 25% quartiles of the dynamicity

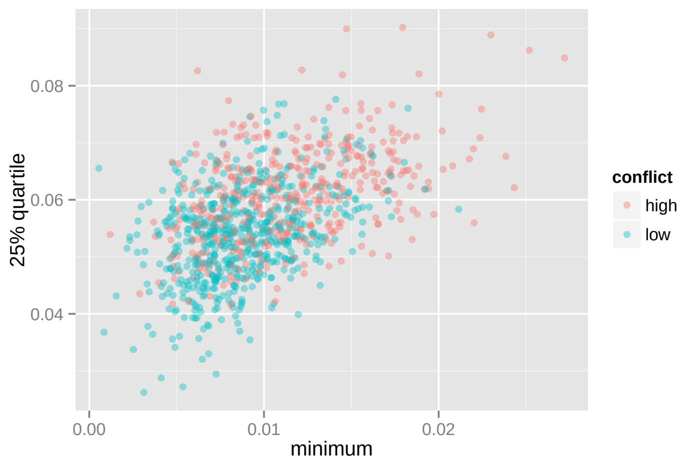


Figure 4.5: Spread of the training and development set of the conflict dataset, with respect to the minimum and the 25% quartile of the dynamicity contour.

contour, than those of the low conflict recordings. Based on the Figure 4.6, we can notice that positive arousal samples tend to have a higher variation and values, in terms of the minimum and 25% quartiles of the contour, than those of the negative arousal samples. And finally, based on the Figure 4.7, in the context of the autism problem, typical speech samples tend to take lower minimum and 75% quartiles of the dynamicity contour, relative to the atypical samples.

4.3 Spectral Emotion Profile

The purpose of Spectral Emotion Profile is to highlight the spectral differences of individuals in expressing affect, and to make use of those differences to personalize the recognition of affective speech. To define spectral emotion profile, we have taken advantage of spectral energy distribution. Validity of the proposed idea is verified in the contexts of discrete and continuous recognition of emotion, using EMO-DB and VAM datasets, respectively. Result of the experimental study show how different spectral intervals of individual speakers, as well as those of different genders, vary in contribution to emotional expression of speech.

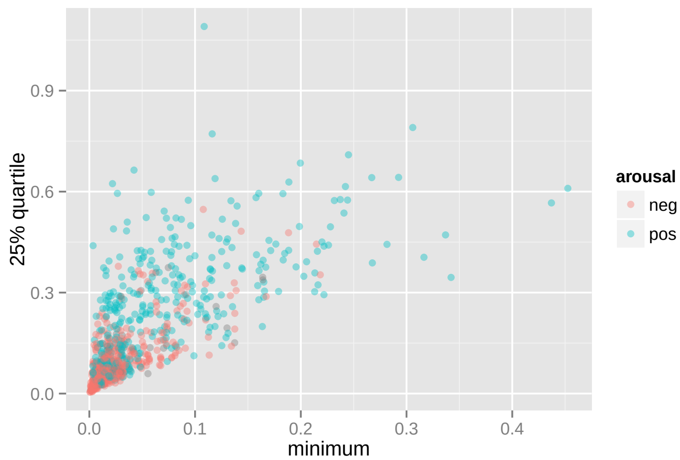


Figure 4.6: Spread of the training and development set of the emotion dataset, with respect to the minimum and the 25% quartile of the dynamicity contour. (gray colored points represent those samples for which the arousal label is missing)

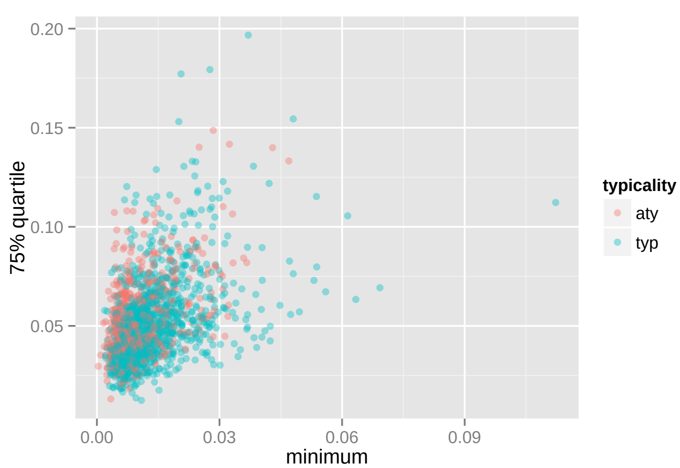


Figure 4.7: Spread of the training and development set of the autism dataset, with respect to the minimum and the 75% quartile of the dynamicity contour.

4.3.1 Background

Asking a good question can be as important as answering the existing questions–if not more so. Unlike most of the works on the modeling aspects of emotional speech recognition, this work does not answer the question of what learning or inference algorithms are most suitable for emotional speech recognition. Nor does it discuss dimensionality reduction aspects [53] of emotional speech recognition. Instead, we ask two tightly related questions, and then try to answer them. Question 1: *Given the identity of a speaker, how can we improve an emotional speech recognition system?* That is, to what extent one can personalize the components of such a system. Prior to this work, some works like that of Grimm and others [77] have targeted the problem of speaker-dependent emotional speech recognition. However, those works do not suggest any adjustment at the feature level, focusing instead on the modeling aspects of a speaker-dependent system. In contrast, the focus of this work is mainly on the personalization of the feature extraction component.

It is discussed that the appearance of emotions in speech (and other modalities) vary according to the personal and cultural backgrounds of an individual [172]. This brings up the Question 2: *How different speakers are different in expressing emotions?* In other words, by asking this question we emphasize those aspects of different speakers that make them distinctive in expressing emotional speech. Having a good answer to the second question can pave the way to a good answer for the first question. We answer these two questions by suggesting the notion of spectral emotion profile. To begin with, we review a set of speech features that are called spectral energy distribution [54, 55]. Accordingly, we show how different spectral intervals in the human voice contribute differently to emotional expressions, and how those intervals can change from one individual to another.

Based on the proposed notion of spectral emotion profile, we have run two experiments. In one experiment, we observe the spectral emotion profile of a group of speakers altogether, alongside with those of the each of the genders from the same group. The purpose of this experiment is to show how different spectral intervals contribute differently in the expression of emotions for the two genders. In another experiment, we observe the spectral emotion profiles of individual speakers, to show the diverse applicability of each spectral interval for different individuals. As for the choice of emotional speech data, we have adopted the VAM and the EMO-DB databases.

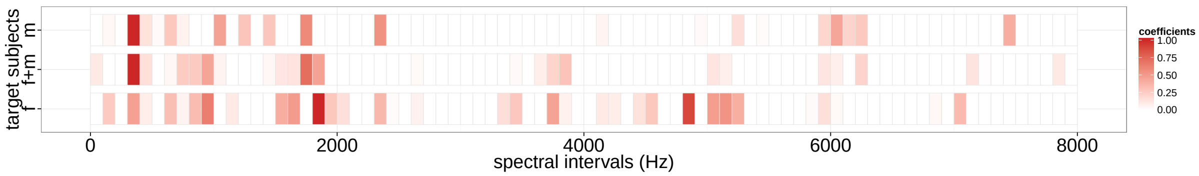
4.3.2 Spectral Emotion Profile

As the name suggests, the spectral emotion profile or SED is proposed to highlight the spectral characteristics of a human voice that impact emotional expressions. Our main intention for this proposition is to enable personalization of speech emotion recognition at a lower level than that of the training stage of a predefined model. That is to say, for such a system, we would like to have personalized model, which also involves a personalized choice of speech features. To mention a few cases of applicability of such system, SEP can be defined for individuals, as well as groups of people who share one or more characteristics. For example, it can be applied to groups of people with the same language, cultural background, or gender. Therefore, either for modeling purposes, in which a model is supposed to target a specific group of subjects, or for subjective studies of emotional speech, with an interest in the characteristics of human subjects and how they can influence the appearance of emotions, SEP can be employed effectively.

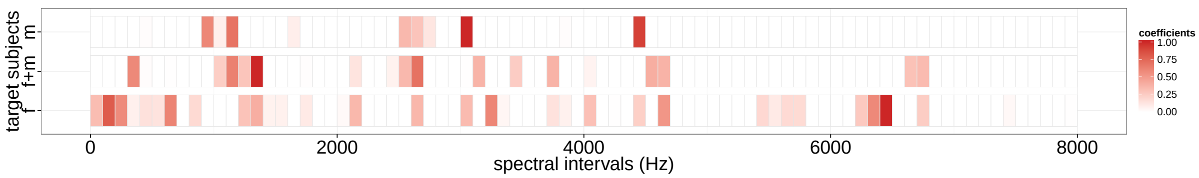
Variable Selection

Based on to the definition of spectral energy distribution, we now define the notion of the spectral emotion profile (SEP) as follows. SEP specifies those SED components, or equivalently spectral intervals, that are relatively more informative than others, when used for capturing the emotional contents of individual or particular groups of speakers. Hence, to continue with the idea of the spectral emotion profile, we need to select a subset of SED components. Selection, as intended here is to find a subset of features that are the most descriptive for the purposes of a study. And additionally, to show the relative importance of each the selected subset of features. For instance, assuming that the aim of a learning task is to distinguish positive and negative valence speech, selection will be performed to find the most informative subset of features that can discriminate positive and negative valence samples.

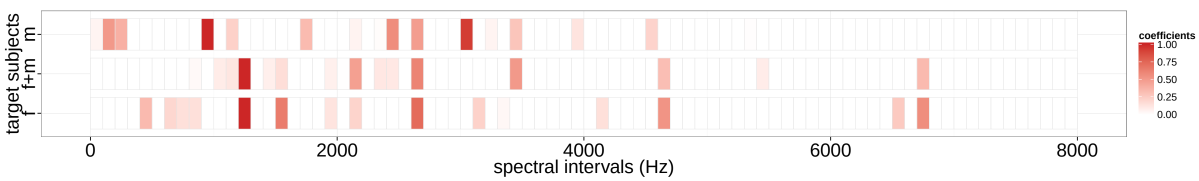
As for the different choices of variable selection algorithms, one can think of correlation-based [80] or information theoretic [148] selection methods, as well as Laplacian [82] and Fisher scores. These sorts of algorithms, known as filters, optimize different objective functions which have to do with some senses of separability of features, however do not necessarily guarantee the optimality of the resulting model, as those senses of separability are not directly relevant to the ultimate purpose of a model that is minimizing the prediction error. To take into account that matter, we are inclined to use shrinkage algorithms [81]. Those algorithms embed the selection into the learning procedure of parameters, therefore carry out selection aligned with the ultimate modeling objectives. Our choice of shrinkage algorithm in this work is the lasso [81], due to the sparsity that it suggest. To achieve that, what lasso does is to penalize those regressors that do not contribute to a better description of the response variables. In the context of SEP, regressors are the SED components and the response variables are the emotional contents of speech. Therefore, what we will gain as a result is a subset of SED components that contribute to the optimality of a model of emotional speech.



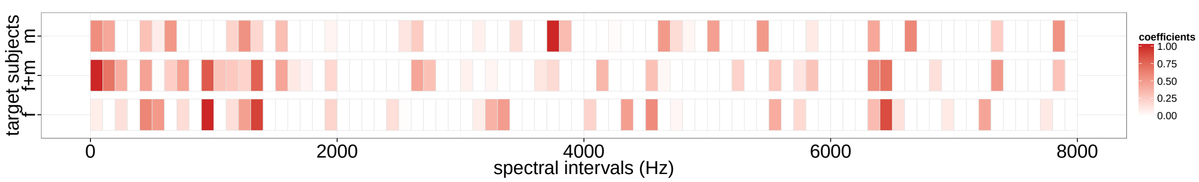
(a) EMO-DB (f: female, m: male)



(b) VAM - Activation Dimension (f: female, m: male)



(c) VAM - Dominance Dimension (f: female, m: male)

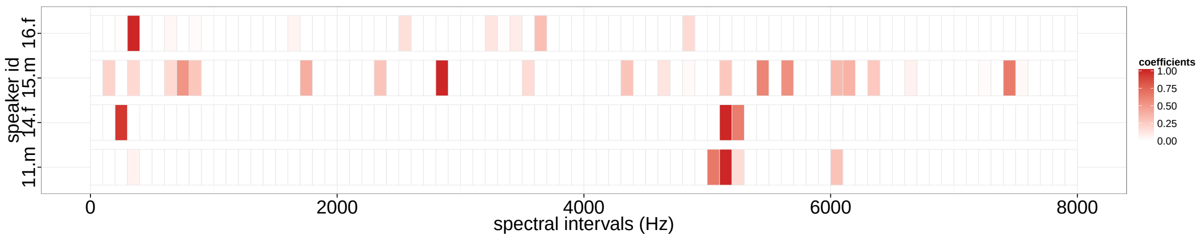


(d) VAM - Valence Dimension (f: female, m: male)

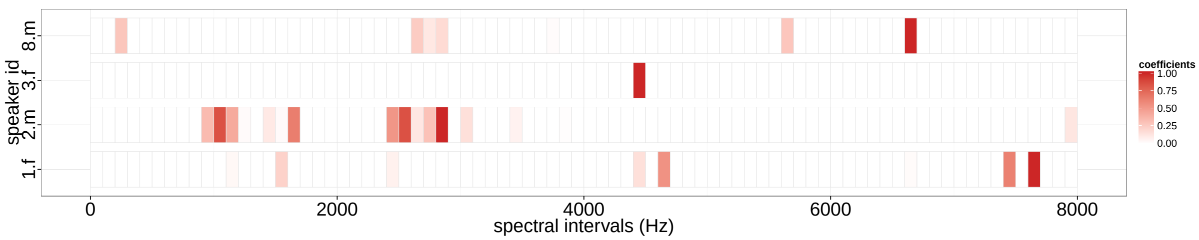
Figure 4.8: Genders Spectral Emotion Profile

Visual Presentation

Given the SEP of different individuals, we would like to compare different SED components, or analogously different spectral intervals, based on their relative importance. For instance, let us assume that in an experiment, all spectral intervals' size is set to 100 Hz and that they cover 0-8 kHz. Therefore, there will be 80 (i.e., $\frac{8000-0}{100}$) SED components, which in turn will result in 80 evaluations of those components. As perceiving those 80 numbers may not be very easily attainable, we use a visual representation of SEP. To do so, we use a heat map of those evaluations, where the map is composed of 80 tiles, that is as many as the number of SED components, and the shade of each tile indicates the relative importance of the corresponding component.



(a) EMO-DB (f: female, m: male)



(b) VAM - Activation (f: female, m: male)

Figure 4.9: Individual Speakers Spectral Emotion Profile

4.3.3 Experimental Study: Genders

Setup

To investigate the dependency of the effective choice of SED components on the speakers' gender, we make use of the spectral emotion profile in this experiment. For this purpose we use the EMO-DB and VAM databases. For the sake of a fair comparison, as the number of recordings by female speakers in the two databases is more than that of the male speakers, we use a subset of each database. The selected subset in either of the two cases is the biggest subset of all the recordings that include an equal number of recordings by female and male speakers.

The objective of this experiment is to find a subset of SED components that maximizes the prediction accuracy of a model that uses that subset as its regressors. For the EMO-DB, we use one model for distinguishing all the emotional labels, whereas for the VAM, we use three different models, one for each of the emotion dimensions. The implemented models are of the family of linear models and as for the learning algorithm we use the lasso. Therefore, to evaluate the relative importance of SED components for each of the learning tasks, we use the magnitude of the regression coefficients. As the magnitude of each component would affect the magnitude of the corresponding regression coefficient, we have normalized each component in advance to the learning phase. Hence, the more the relative magnitude of a coefficient, the higher the importance of the corresponding SED component is (for comparison purposes, the normalized absolute values of the regression coeffieicnts are used). Due to the sparsity of the lasso solutions, a good number of coefficients will be zero, which is a good thing, as we will end up with a more concise model. As for the parameter setting, we use 10-fold cross validation.

SED components are extracted from the range 0-8 KHz (restricted by the sampling frequency of the recordings – Nyquist theory) by setting the spectral length of each interval to 100 Hz. As for the choice of the exponent p in equation 4.4 for the EMO-DB, this value is set to 0.2, and for the VAM to 0.2, 0.3, and 0.15, for each of the activation, dominance, and valence dimensions, respectively. These parameters are set according to previous works [53–55], for which they were optimized.

Results and Discussion

A visualization of the resulting profiles are shown in Figure 4.8. For each database, we can notice the difference between the selected spectral intervals for each gender, and how each profile shows a limited number of SED components to be useful for each of the modeling tasks. Based on this experiment, it is evident how ignoring the gender of a subject (`f+m`) can make a noticeable difference in the choice of the speech features, compared to the cases where the gender has been take into account (`f` and `m`). Although there are some overlaps for the choice of SED components for each of these three cases, the difference between one to another is obvious.

4.3.4 Experimental Study: Individual Speakers

Setup

By means of This experiment, we investigate the application of spectral emotion profile to individual speakers, that is the dependency of an effective choice of SED components on the identity of speakers. For this experiment too we use the EMO-DB and VAM databases. Again, for the sake of the fairness of the comparisons, we use a subset of the dataset, comprising four speaker, two female and two male speakers, where the number of recordings for each speaker is equalized.

The objective of this experiment is to investigate the dependency of an efficient set of features on the individual speakers. In this experiment too we use the family of linear models as for the choice of model, and the lasso as for the learning algorithm. The objective of each learning task is to train a model for each of the actors, so that the model can maximize the prediction accuracy of the samples from the corresponding individual speaker. For the VAM database, we are only using the activation dimension in this experiment, as that can sufficiently deliver the objective of this experiment. Similar to the previous experiment, the magnitude of the regression coefficients are used to compare the relative importance of each SED component for each of the modeling tasks. Also similar to the previous experiment, features are normalized prior to the learning phases, and 10-fold cross validation is used to set model parameters.

SED components are extracted from the range 0-8 KHz (restricted by the sampling frequency of the recordings – Nyquist theory) by setting the spectral length of each interval to 100 Hz. As for the choice of the exponent p in equation 4.4 for the EMO-DB, this value is set to 0.2, and for the VAM to 0.2, 0.3, and 0.15, for each of the activation, dominance, and valence dimensions, respectively. These parameters are set according to previous works [53–55], for which they were optimized.

Results and Discussion

The resulting individuals' spectral emotion profiles are visualized in Figure 4.9. Among the first things that one may notice from this figure is the sparsity of the profiles compared to the previous experiment. This observation suggests that the complexity of a model for an individual's emotion recognition can be much lower, compared to that of a model that is used for a general cases, regardless of the target subject. In addition, by cross comparing different individuals, we can see how the selected SED component for each individual is different than all the others.

Presented in this work is the notion of spectral emotion profile (SEP), as an answer to two questions: 1) what are the ways to personalize an emotional speech recognition system at a feature level? And 2) what makes different individuals different in expressing emotions through speech? To define SEP, we used a set of spectral energy distribution (SED) and defined SEP as a subset of SED components, or analogously spectral intervals, that contribute to the emotional statement of a piece of speech. Through the idea of spectral emotion profiles, and by the use of EMO-DB and VAM emotional speech databases, we observed how the gender and identity of target subjects can play a key role in an optimal choice of speech features. This information can be effectively used for the personalization of speech emotion recognition system.

4.4 Conclusion

Primarily proposed in this work is spectral energy distribution (SED) as a set of low-level descriptors. Capabilities of SED is shown through a set of experiments that include continuous and binary dimensional recognition of affect, as well as recognition of trait from human speech. Furthermore, based on SED, a concept of dynamicity is proposed that measures the amount of temporary changes in speech signal. This is put into practice by using it to recognize affect, autism, and conflict from speech. Finally, the concept of spectral emotion profile (SEP) is proposed that captures context-dependence differences in conveying affect that are reflected in the spectrum of their speech. Validity of SEP is verified by using it to model individual as well as gender-dependent affective profiles for modeling categorical and dimensional affect.

Chapter 5

Proposed Statistical Modeling Solutions

5.1 Variable Selection

Number of speech features that were commonly used for recognition of affect from speech a few years ago exceeded a thousand. Today, this number exceeds six thousands. However, not all of those features are useful for extracting affective contents of speech. On the one hand, extracting features that are not relevant, and learning a model based on such features could be a computational overcharge. On the other hand, irrelevant features can degrade generalization capabilities of an estimation. Therefore, variable selection becomes an imperative part of a solution for affective speech recognition. In this section, we examine the effectiveness of different dimensionality reduction algorithms for this purpose. The selected set of algorithms include projection methods, as well as filters and embedded methods. Our experiments are based on the VAM affective speech dataset, as well as Interspeech 2012 and 2013 paralinguistic challenges. In this study, we distinguish between low-level descriptors and functionals as the two components of a speech features, and we study the relevance of each to capturing the affective information.

5.1.1 Background

A solution to paralinguistic speech recognition is essentially composed of two main stages: signal processing, and statistical modeling. At the signal processing stage, extraction of speech features is concerned. At the statistical modeling stage, however, the objective is to find a mathematical relationship between those features and paralinguistic qualities. The common practice for tackling this problem is to extract a long list of features from speech, as long as a few thousands [176, 184–189], and to fit a model that optimizes a sense of fitness, accordingly. What is of vital necessity in this procedure is feature selection. That is, to identify a set of speech features that are useful for a particular recognition task, and to remove the rest. There are several advantages to performing feature selection. One is that the reduced list of features can be used to provide explanation as to how paralinguistic qualities are related to physical speech phenomena. Such explanations are particularly useful for simulating speech samples that are parlinguistically modified, e.g., a system that synthesizes affective speech.

From another point of view, the more features we use to build a statistical model, the more parameters such a model would have, which in turn implies higher degree of complexity. A model with a higher degree of complexity is deemed to suffer more from over-fitting. Therefore, the second advantage of feature selection is that the resulting model is more likely to be able to generalize well.

From a third point of view, a model that uses more speech features is considered to be computationally more expensive. On the one hand, more number of features requires more expensive feature extraction. And, since features are extracted from short time-frames, e.g., from 20 millisecond long frames, having a list of some thousands features implies extracting those from each frame, which explains the expensive cost of feature extraction. On the other hand, training and recall times of a statistical model are directly affected by the length of the features vector. In spite of very advanced processor technology nowadays, there are still cases for which we need to carefully spend the processing resources. An example of such cases is real-time applications, specially for mobile platforms, where efficiently allocating resources is of yet higher importance. Therefore, besides the potential benefits that feature selection may have from the accuracy point of view, it has a distinct advantage and that is the fact that in its absence real-time mobile application of paralinguistic speech recognition might be inaccessible.

The objective of this study is therefore to investigate the choice of features for recognition of different paralinguistic qualities from speech: affect, autism, conflict, likability, pathology, and personality. Results of this study show that by carefully selecting a subset of features, we can achieve higher prediction accuracy, and the same time extract far less number of features. In the variable selection that we perform, we distinguish between selected set of low-level descriptors and the selected set of functionals.

As follows, we discuss different approaches to dimensionality reduction, and we present our rationale for choosing variable selection for reducing the dimensionality. This is followed by an introduction to the variable selection algorithm of choice in this work. In the first experiment, we compare four different dimensionality reduction algorithms with the objective of learning for modeling continuous affect. Then, we present the results of variable selection for 11 different paralinguistic qualities of speech, by discussing the chosen set of low-level descriptors, as well as the choice of functionals, and comparing the prediction accuracy in the absence of variable selection to the results of our study.

5.1.2 Dimensionality Reduction

Given the explanatory variables $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\} = \mathbf{X} \in \mathcal{X}$ and the response variable $\mathbf{y}$, the objective of dimensionality reduction is to find a subspace $\hat{\mathcal{X}} \subseteq \mathcal{X}$ with minimal dimensionality d that can satisfy a particular criterion. In the case of supervised learning, the criterion is to maximize the prediction accuracy.

Regardless of the nature of a learning problem, which can be either supervised or unsupervised, in search for the subspace $\hat{\mathcal{X}}$, either to take the response variable $\mathbf{y}$ into consideration or not, is the matter of supervised or unsupervised dimensionality reduction. Since we are interested in finding a set of covariates that are most effective for paralinguistic speech recognition, we focus on supervised algorithms.

From another point of view, dimensionality reduction algorithms can be grouped into two categories: projection and variable selection. While projection algorithms redefine covariates by suggesting a combination of the original ones, selection algorithms reduce the set of covariates to a subset of those. The two categories can be unified by the following equation:

$$\hat{\mathbf{X}} = \mathbf{X}\mathbf{W}, \tag{5.1}$$

where $\mathbf{X}$ is $N \times p$, with N and p being the number of samples and the original number of covariates, and $\mathbf{W}$ is $p \times d$, where d is the dimensionality of the destination space. For projection algorithms, entries of $\mathbf{W}$ may take values from all over the range. However, for selection algorithms, entries of $\mathbf{W}$ take values from $\{0, 1\}$. Additionally, for selection algorithms, each column of the matrix has only one none zero entry, where the index of that entry indicates the index of a selected variable.

Depending on one's perspective, one of the two could be preferable. On the one hand, selection algorithms preserve the nature of the original covariates, therefore give way to an interpretable model. Moreover, for real-time application of paralinguistic speech recognition, another advantage of selection algorithms is that they remove the necessity of extracting all the original features from speech, which in turn may result in a far smaller computational expense. On the other hand, the sparsity condition on the transformation matrix $\mathbf{W}$, may add to the computational complexity of selection methods.

For the variable selection category there are two further sub-categories: wrappers and filters. The difference between the two is in the decision criterion that each uses to select a subset of variables. Where wrappers depend on the prediction error, filters use some criteria independent of

Table 5.1: Summary of dimensionality reduction results (CC: correlation coefficient, MLE: mean absolute error, NF: number of features)

Feature Set	Reduction Method	Valence CC(MAE)	Valence NF	Activation CC(MAE)	Activation NF	Dominance CC(MAE)	Dominance NF	**mean** CC(MAE)	**mean** NF
a+b	PCA	0.42(0.14)	148	0.82(0.16)	**60**	0.80(0.14)	**38**	0.68(0.15)	82
	Greedy FS	0.42(0.14)	**86**	0.82(0.16)	78	0.80(0.14)	102	0.68(0.15)	89
	Elastic Net	**0.44(0.13)**	191	**0.83(0.15)**	156	**0.81(0.14)**	199	**0.69(0.14)**	182
	SPCA	0.42(0.14)	114	0.82(0.15)	**52**	0.80(0.14)	**42**	0.68(0.14)	**69**
a	PCA	0.38(0.14)	88	0.81(0.16)	**38**	0.80(0.14)	**26**	0.66(0.15)	51
	Greedy FS	0.37(0.14)	**42**	0.80(0.16)	52	0.78(0.17)	70	0.65(0.16)	55
	Elastic Net	**0.39(0.14)**	167	0.81(0.16)	92	**0.81(0.14)**	145	**0.67(0.15)**	135
	SPCA	0.38(0.14)	**36**	0.81(0.16)	**34**	0.80(0.14)	34	0.66(0.15)	**34**
b	PCA	**0.42(0.14)**	64	0.79(0.17)	62	0.76(0.15)	**36**	**0.66(0.15)**	54
	Greedy FS	0.41(0.14)	**48**	0.79(0.17)	**48**	0.76(0.15)	52	0.65(0.15)	**49**
	Elastic Net	0.41(0.14)	78	0.79(0.17)	**49**	0.76(0.15)	45	0.65(0.15)	57
	SPCA	0.40(0.14)	70	0.79(0.17)	56	0.76(0.15)	52	0.65(0.15)	59

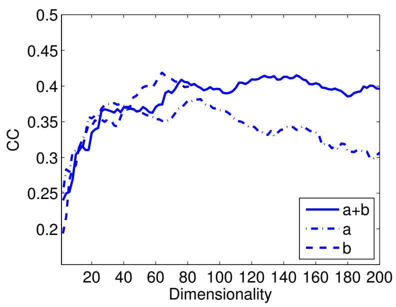
the prediction result, although somehow relevant. Despite the fact that wrappers can potentially converge to the globally optimal subset of variables, they could be computationally expensive, considering that in order to exhaust all the possibilities, 2^p different subsets should be seen. On the contrary, filter methods are not as computationally demanding, however due to the departure of selection criteria from the prediction error, suboptimality of a solution is not unexpected.

5.1.3 Experimental Study: Comparing Dimensionality Reduction Algorithms

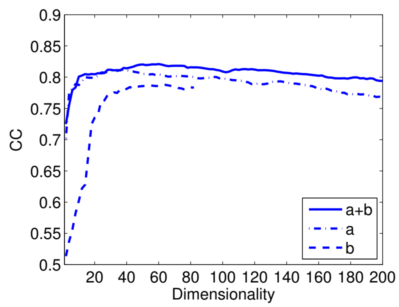
Setup

We use the VAM dataset throughout these experiments. Therefore, the objective is to reduced the dimensionality of the feature vector with the objective of learning a regression model for modeling continuous affect. We use the linear model for regression, and we estimate the coefficient of the linear model using maximum-likelihood estimation, assuming Gaussian distribution for the prediction error.

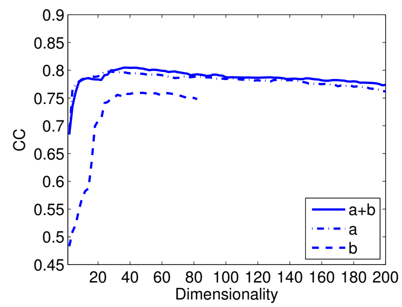
We have extracted two sets of features in this work.



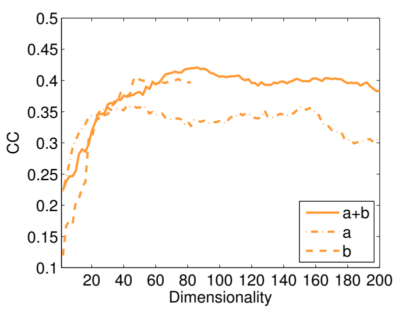
(a) PCA - Valence



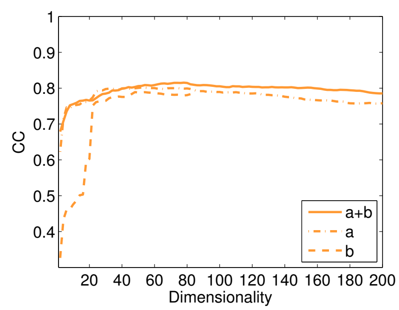
(b) PCA - Activation



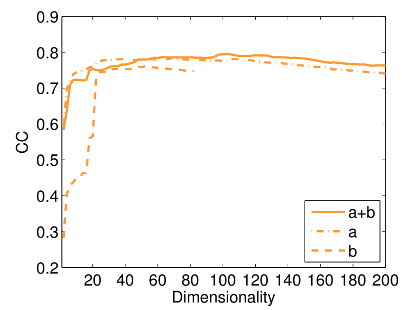
(c) PCA - Dominance



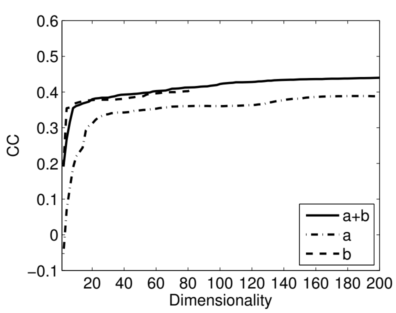
(d) Greedy FS - Valence



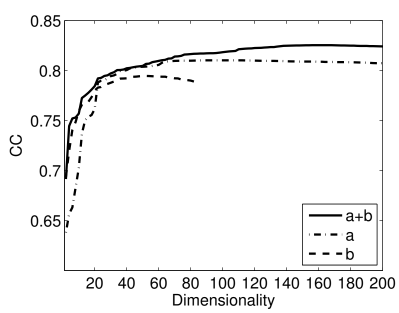
(e) Greedy FS - Activation



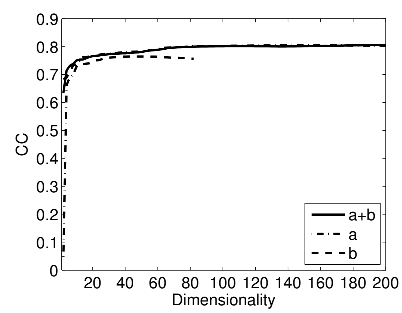
(f) Greedy FS - Dominance



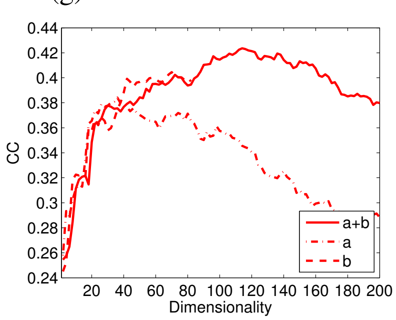
(g) Elastic Net - Valence



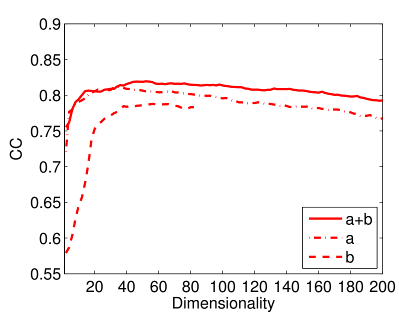
(h) Elastic Net - Activation



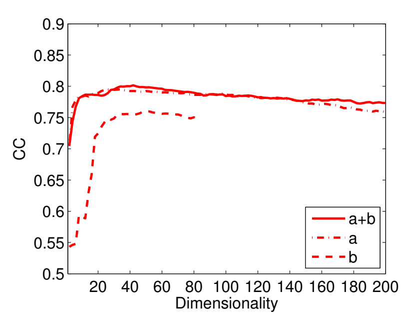
(i) Elastic Net - Dominance



(j) SPCA - Valence

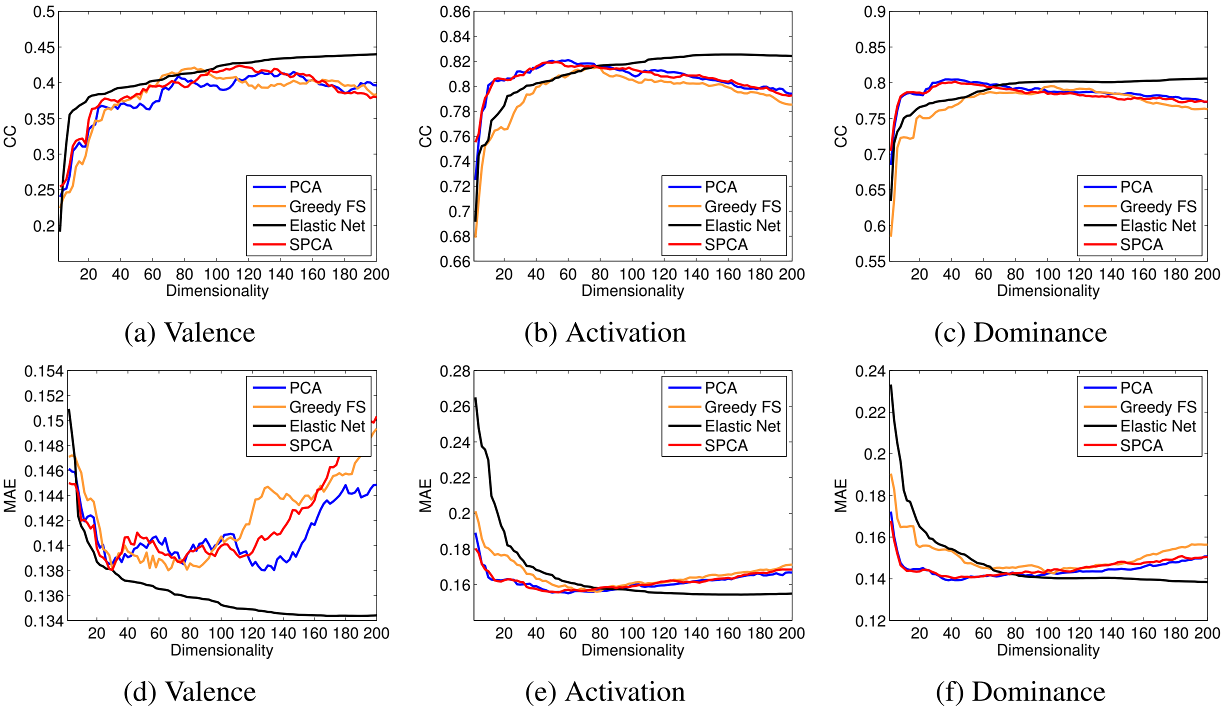


(k) SPCA - Activation



(l) SPCA - Dominance

Figure 5.1: Dimensionality reduction results (CC: correlation coefficient)



(a) Valence (b) Activation (c) Dominance

(d) Valence (e) Activation (f) Dominance

Figure 5.2: Dimensionality reduction results for the `a+b` feature set (CC: correlation coefficient, MLE: mean absolute error)

set a. This is composed of a set of features that we call spectral energy distribution (SED). In this work, we have set the parameters so that $l_1 = 0Hz$, $l_i = u_{i-1}$, and $u_i - l_i = 100Hz$. The components cover the whole spectral range (0-8KHz according to the database of interest). The function $g(.)$ (Eq. 4.2 and 4.4) is set to the family of rational exponents, with exponents of 0.15, 0.2, and 0.3 for each one of the valence, activation, and dominance dimensions. SED is extracted from 100 msec windows, and as for the statistics, we have used the minimum, maximum, mean, median, and the variance. The total number of features in this set is 400.

set b. The features in this set are the fundamental frequency, the first three formants, the first twelve MFCCs, total energy, and the zero cross-over rate. For the fundamental frequency, formants, and MFCCs extraction is done from 50 mSec windows of signal; we have then computed the minimum, maximum, mean, median, and variance. Number of features in this set add up to 82.

For experimental purposes, we use each of the two feature sets a and b, as well as their combination, which we denote by a+b.

10-fold cross validation (CV) is set as a standard [92, 173] for evaluating prediction result on the VAM database and we adopt to that. For the sake of fairness of the comparisons that we are going to make among the four algorithms, we will fix the CV indices throughout the experiment. Pearson's correlation coefficient (CC) and mean absolute error (MAE, also referred to as mean linear error or MLE) has been used as the means of evaluation of the prediction accuracy.

Results and Discussion

The result of experiments is shown on Figures 5.1 and 5.2, and Table 5.1. According to Figure 5.1, we can see that, regardless of the choice of dimensionality reduction method, the feature set a+b, although contains all the features from feature sets a and b, does not necessarily lead to a better prediction accuracy, compared to the scenarios that we used just one of the two. In other words, from an optimization point of view, all the four dimensionality reduction methods that we used in this study are likely to suffer from sub-optimality. In general, we can see that for the relatively smaller number of features, usually the feature sets a and b result in a better prediction accuracy than that of the a+b. Now, let us take a look at a few different cases of sub-optimality of solutions, based on the results presented on Figure 5.1.

- According to the Figure 5.1b, the best accuracy obtained resulting from dimensionality reduction by PCA belongs to the feature set b. This means that regardless of the dimensionality, PCA could not find a transformation of the `a+b` space that can result in a prediction as accurate as that of the `b` space.

- For the valence dimension and dimensionalities less than 60, both greedy feature selection and SPCA's most accurate prediction is obtained by one of the feature sets `a` or `b`.
- When we used the elastic net for the dominance dimension (Figure 5.1i), the accuracy of the prediction using the feature set `a`, although not significantly, outperforms that of the feature set `a+b`, for the most part of the dimensionality range.

Now, we would like to compare the prediction accuracy resulting from dimensionality reduction by the four algorithms, for each of the emotional primitives (Table 5.1). For the valence dimension, for feature sets `a` and `a+b`, elastic net's reduction leads to the best accuracy, although it takes the highest number of dimensions among all to obtain that accuracy. For the same feature sets, the accuracy of the prediction resulting from reduction by the greedy feature selection is comparable to that of the elastic net, but it takes the least dimensionality for the greedy feature selection to do the job. For feature set `b`, however, PCA's reduction gives the most accurate prediction among all the four algorithms. For the activation dimension, for the feature set `a+b`, elastic net's reduction results in the most accurate prediction, which again comes at the price of the highest number of dimensions. For the same dimension and feature set, SPCA results in a comparable accuracy to that of the elastic net, however with way less number of dimensions (33%). For the dominance emotion primitive, for the feature set `a+b`, elastic net's reduction results in the most accurate prediction, and again with the highest number of dimensions. For the same dimension and feature set, both PCA and SPCA result in comparable prediction accuracies to that of the elastic net, however with significantly less number of dimensions (about 20%).

To see where the results of this work stand compared to recent works on the same database, Table 5.2 puts two sets of results of this work side-by-side with those of a work by Schuller [173]. The first set is the one which has the most accurate predictions (elastic net, feature set `a+b`) and the second set has the next most accurate predictions and at the same time the least number of features (SPCA, feature set `a+b`). Based on the results presented on this table, we can see that the average prediction accuracy of the elastic net is higher than the other two, however the dimensionality of the feature vector used for this task is considerably greater than those. On the other hand, SPCA offers a relatively accurate prediction accuracy, but the shortest feature vector, compared to the other two.

Table 5.2: A comparison (CC: correlation coefficient, MLE: mean absolute error, NF: number of features)

	Valence		Activation		Dominance		**mean**	
	CC(MAE)	NF	CC(MAE)	NF	CC(MAE)	NF	CC(MAE)	NF
Schuller [173]	**0.45(0.13)**	238	0.81(0.16)	109	0.79(0.14)	88	0.68(0.14)	145
this work (elastic net)	0.44(0.13)	191	**0.83(0.15)**	156	**0.81(0.14)**	199	**0.69(0.14)**	182
this work (SPCA)	0.42(0.14)	**114**	0.82(0.15)	**52**	0.80(0.14)	**42**	0.68(0.14)	**69**

5.1.4 Experimental Study: Variable Selection

Setup

Datasets that are considered for this study are used as benchmark of Interspeech paralinguistic challenges, therefore they come in train and test subsets. For each selection task that we discuss, selection is performed based on the training subset of the data, by evaluating subsets of features using 5-fold cross validation. The selected subset of features for each paralinguistic quality is the one that results in the highest cross validation accuracy. For each selection task, we also report prediction accuracy for the test set, using the model that uses the selected subset of features. Granularity of interest in all these experiments is sentence-long. As for the selection algorithm, we use the randomized generalized linear models (RGLM).

Selection method: Random Generalized Linear Model

Random generalized linear model or RGLM [195] is an ensemble learning algorithm based on bootstrap aggregation of generalized linear models. Steps of the algorithm are as follows:

1. A number of bootstrap subsets are chosen by sampling with replacement from the training set. Each subset is called a bag.

2. A set of features is randomly chosen for each bag.

3. The set of selected features in each bag are ranked based on their relevance to the output.

4. Top-ranked features of each bag are used for forward selection. Akaike information criterion (AIC) is used as the selection criterion.

5. For each bag a generalized linear model is fit.

The result of this algorithm is a number of generalized linear models that each uses a different set of features, and each is constructed based on a different subset of training samples. The randomness of the RGLM is due to these two sources of randomness. Later on, to recall the model, output of all models across bags are aggregated to arrive at a final output. Selected set of features by the model is the aggregation of all those selected across all bags.

Affect

Categorizing affective speech according to the arousal and valence dimensions is considered in this experiment, and the Geneva multi-modal emotional portrayals dataset is used for this purpose.

Selected sets of LLDs and functionals for arousal are shown in Figures 5.3a and 5.3b, respectively. According to Figure 5.3a, Rasta-filtered spectrum LLDs comprise the majority of the selected features. Moreover, lower and upper MFCC coefficients, spectral roll-off, fundamental frequency, and harmonicity show to be useful for modeling arousal. According to Figure 5.3b, static functionals are more useful than dynamic ones. Among the static functionals, location statistics are shown to be more descriptive than spread statistics. Among the location statistics, first percentile, first quartile, and root quadratic mean are the most useful. On the other hand, flatness and standard deviation are among the most informative spread functionals. As for the dynamic functionals, linear predictive coefficients are shown to be by far the most descriptive dynamic functional.

For the arousal dimension, the reduced set of features comprises 41 LLDs (62%) and 30 functionals (54%), which adds up to 226 features. This is less than 4% of the baseline features. This has resulted in a prediction accuracy of 78.5%, which is lower by 3.9% than the case where we use all the 6373 baseline features.

Selected sets of LLDs and functionals for valence are shown in Figures 5.3c and 5.3d, respectively. Selected set of LLDs for the valence dimension show to be more spread among different types. Although Rasta-filtered spectrum and MFCC comprise a majority of the selected features, spectral roll-off does not seem to be be as informative for valence as it is for arousal. Instead, spectral flux, entropy, and centroid seem to be of higher importance for modeling valence. Energy and voicing related LLDs have shown to be more essential for modeling valence than for modeling arousal. Among the functionals, unlike arousal, dynamic functionals are more relevant for modeling valence than their static counterparts. Among those functional, quadratic regression coefficients, and the positions of the maximum and minimum are the selected more than any other functional. Among the spread static functionals, flatness, kurtosis, and skewness are the most descriptive. And, among the location statistics, up-level time, first percentile, and quartiles are shown to be more relevant than the rest.

For the valence dimension, on the other hand, the reduced set of features comprises 51 LLDs (77%) and 44 functionals (79%), which adds up to 151 features. This is slightly more than 2% of the baseline features. This has resulted in a prediction accuracy of 72.6%, which is lower by 5.3% than the case where we use all the 6373 baseline features.

Autism

Modeling autism based on speech samples is considered in this experiment. The purpose of the study is to distinguish typical vs atypical development of language in early ages, and the child pathological speech dataset is used for this purpose.

Selected sets of LLDs and functionals for autism are shown in Figures 5.4a and 5.4b, respectively. According to Figure 5.4a, energy related LLDs are relatively more imperative to modeling autism than spectral and voicing related LLDs. Among spectral LLDs, low and low-mid band Rasta-filtered spectrum, low-mid MFCC coefficients, harmonicity, and spectral flux show to be most relevant. According to Figure 5.4b, among different types of functionals, dynamic functionals show to be more relevant than static functionals. Among dynamic functionals, linear predictive coefficients comprise almost all of the selected functionals from that category. Among static functionals, spread statistics seem to be more relevant than location ones. Among spread statistics, flatness, percentile range, range, and standard deviation show to be more relevant that the rest. First and last percentiles are the most selected location statistics.

The reduced set of features comprises 40 LLDs (59%) and 31 functionals (55%), which adds up to 237 features in total. This is 4% of baseline features. This has resulted in a prediction accuracy of 91.3%, which is lower by 1.5% than the case where we use all the 6373 baseline features.

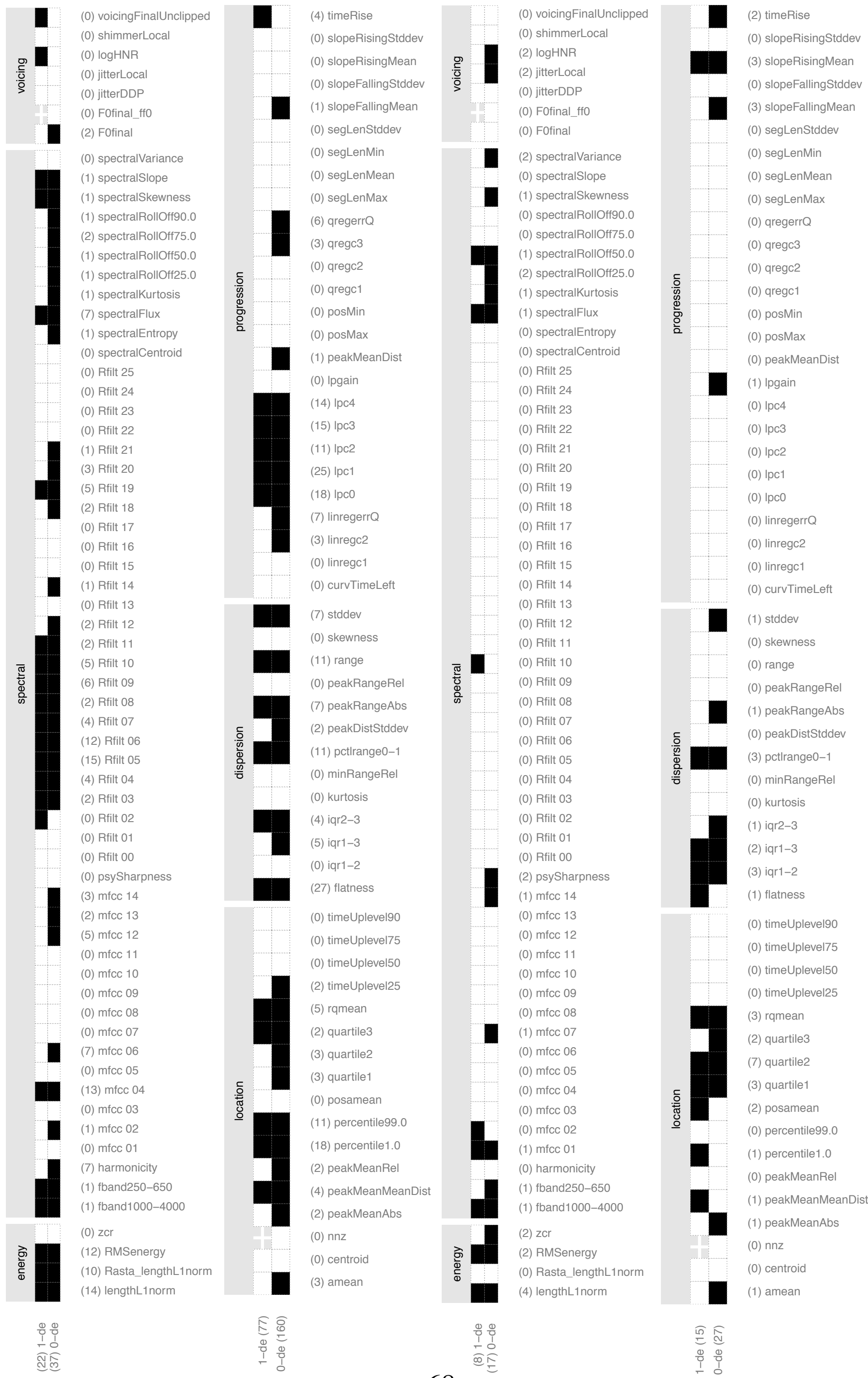
Conflict

Recognition of conflict from speech is the focus of this experiment. For this purpose, the SSPNet conflict corpus is employed.

Selected sets of LLDs and functionals for conflict are shown in Figures 5.4c and 5.4d, respectively. According to Figure 5.4c, energy related LLDs are relatively more relevant to modeling conflict than spectral and voicing related LLDs. Among the spectral LLDs, share of the lower spectral intervals, as well as spectral flux, and roll-off are more relevant than the rest. According

(a) LLD (b) Fcn (c) LLD (d) Fcn

Figure 5.3: Variable selection for arousal (a and b) and valence (c and d). LLD and Fcn stand for low-level descriptor and functional, respectively. (The number in front of each label indicates number of selected features in that category.)



(a) LLD (b) Fcn (c) LLD (d) Fcn

Figure 5.4: Variable selection for autism (a and b) and conflict (c and d). LLD and Fcn stand for low-level descriptor and functional, respectively. (The number in front of each label indicates number of selected features in that category.)

to Figure 5.4d, static functionals are more relevant to modeling conflict than dynamic functionals. Among static functionals, location statistics are more frequently chosen than their spread type counterparts. Among location statistics, quartiles are more relevant than the rest. And, among spread statistics, inter-quartile ranges, and percentile range are more relevant than the rest. Among dynamic functionals, characteristics of slope show to be more frequently chosen than other types.

The reduced set of features comprises 41 LLDs (63%) and 32 functionals (52%), which adds up to 42 features. This is 0.7% of baseline features. This has resulted in a prediction accuracy of 77.6%, which is lower by 1.5% than the case where we use all the 6373 baseline features.

Likability

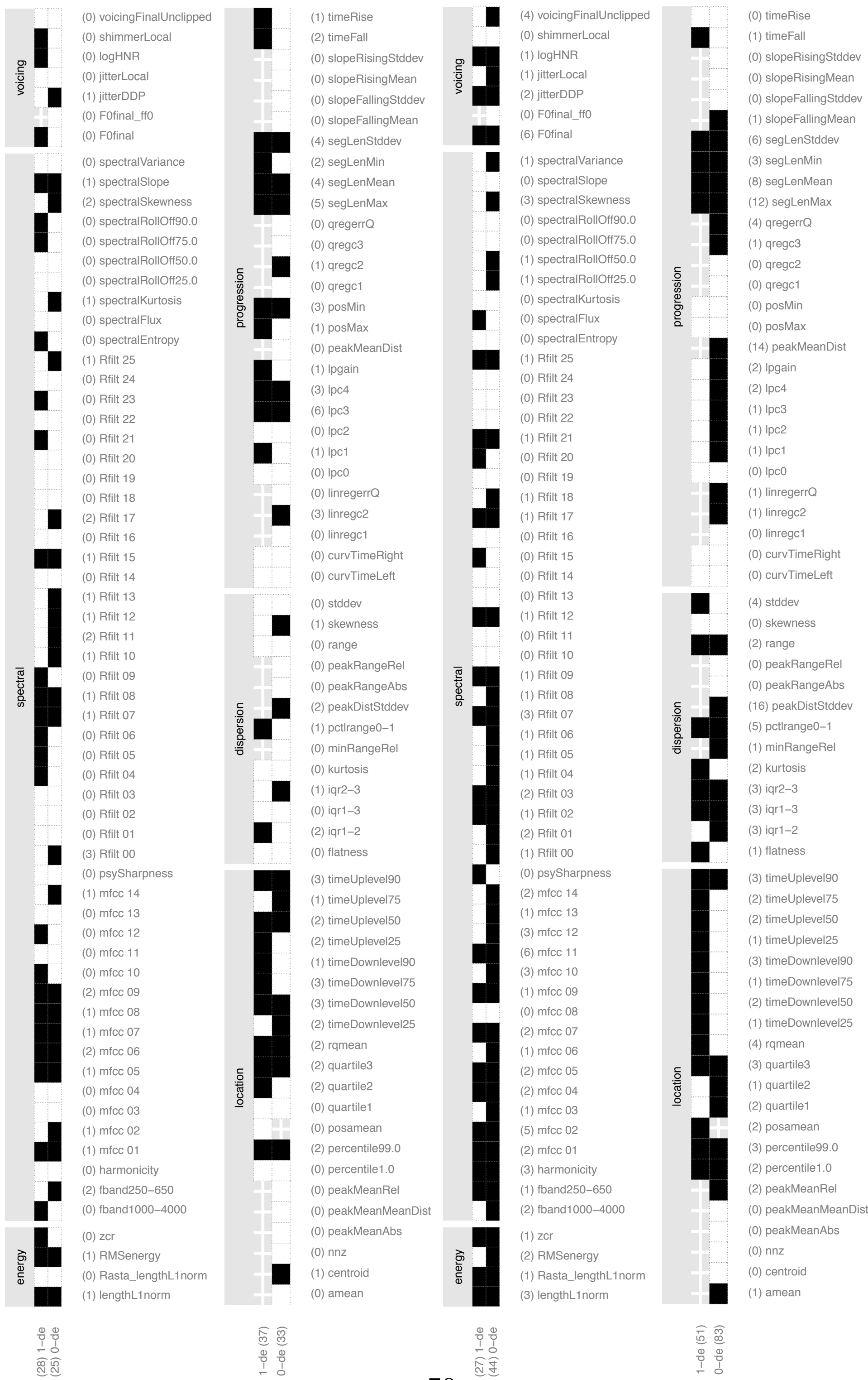
Modeling speaker's likability based on their tone is of interest in this experiment, and the speaker likability database is used.

Selected sets of LLDs and functionals for likability are shown in Figures 5.5a and 5.5b, respectively. According to Figure 5.5a, all three categories of LLDs contribute to the selected set of features more or less to the same amount. Among the spectral LLDs, mid range MFCC coefficients, and mid and low-mid bands of the Rasta-filtered spectrum comprise a major portion of the selected features. According to Figure 5.5b, dynamic functionals, including first-order difference, are more imperative to modeling likability than static functionals. Among other dynamic functionals, segment length, linear predictive coefficients, and the position of peaks are most frequently selected. Among the static functionals, location statistics show to be more relevant than the spread statistics. Among location statistics, down and up-level times, and quartiles are more frequently selected than the rest. Among spread statistics, standard deviation of the distribution of peaks, as well as inter-quartile ranges are among the more relevant choices for modeling likability.

The suggested set of features comprises 41 LLDs (63%) and 32 functionals (52%), which adds up to 70 features. This is almost 1% of baseline features. This has resulted in a prediction accuracy of 60.1%, which is higher by 1.6% than the case where we use all the 6125 baseline features.

Pathology

Recognizing intelligible from non-intelligible status of patients under chemotherapy according to their speech is the focus of this experiment. For this purpose, the NKI speech corpus is used.



(a) LLD (b) Fcn (c) LLD (d) Fcn

Figure 5.5: Variable selection for likability (a and b) and intelligibility (c and d). LLD and Fcn stand for low-level descriptor and functional, respectively. (The number in front of each label indicates number of selected features in that category.)

Selected sets of LLDs and functionals for intelligibility are shown in Figures 5.5c and 5.5d, respectively. According to Figure 5.5c, spectral, energy, and voicing related is the order in which different categories of LLDs contribute to modeling intelligibility. Among spectral intervals, MFCC coefficients, and low and low-mid range Rasta-filtered spectrum, as well as harmonicity and frequency bands are most frequently chosen. Furthermore, fundamental frequency and jitter show to be most imperative to modeling intelligibility, among other voicing related LLDs. According to Figure 5.5d, dynamic functionals show to be more relevant than their static counterparts. Among dynamic functionals, mean of distribution of peaks, and maximum length of segments are selected more frequently. On the other hand, among static functionals, spread statistics show to be more relevant than location statistics. Among spread statistics, standard deviation of the distribution of peaks, and among location statistics, up and down-level times, as well as first and last percentiles are the most frequently used.

The reduced set of features comprises 48 LLDs (74%) and 43 functionals (69%), which adds up to 134 features. This is almost 2% of baseline features. This has resulted in a prediction accuracy of 61.6%, which is higher by 0.5% than the case where we use all the 6125 baseline features.

Personality

Recognizing personality from speech is the focus of this experiment. Five personality dimensions are used for this purpose: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The benchmark of this study is the speaker personality corpus.

Selected sets of LLDs and functionals for openness are shown in Figures 5.6a and 5.6b, respectively. According to Figure 5.6a, voicing related LLDs show to be relatively more frequently selected than spectral or energy related LLDs. Among voicing related descriptors, jitter shows to be very informative for modeling openness. Among spectral descriptors, spectral roll-off, and some sporadic Rasta-filtered spectrum and MFCCs have shown to be useful. According to Figure 5.6b, dynamic functionals, including first-degree difference, show to be more informative than static functionals. Among dynamic functionals, segment length and linear prediction coefficients are more frequently chosen. On the other hand, spread statistics comprise most of the selected functionals from static category. Among the selected functionals from this category, inter-quartile ranges, standard deviation, and kurtosis are the most frequently selected functionals. Among location statistics, up-level time and quartiles show to be more informative for modeling openness.

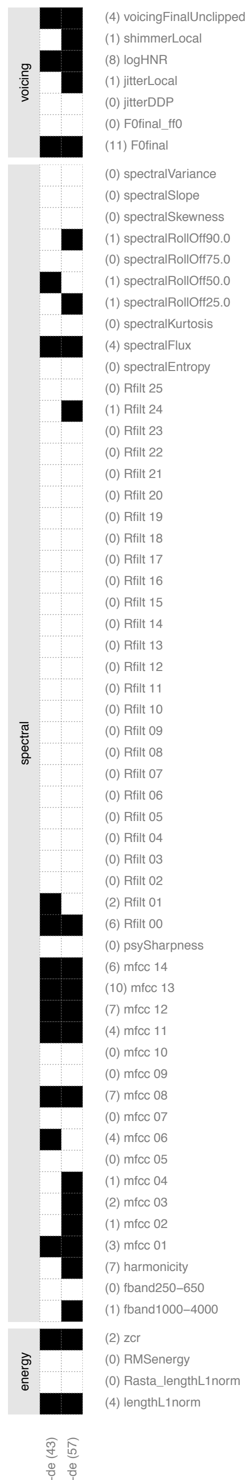
For the openness dimension, the reduced set of features comprises 40 LLDs (62%) and 39 functionals (54%), which adds up to 67 features. This is almost 1% of baseline features. This

(a) LLD (b) Fcn (c) LLD (d) Fcn

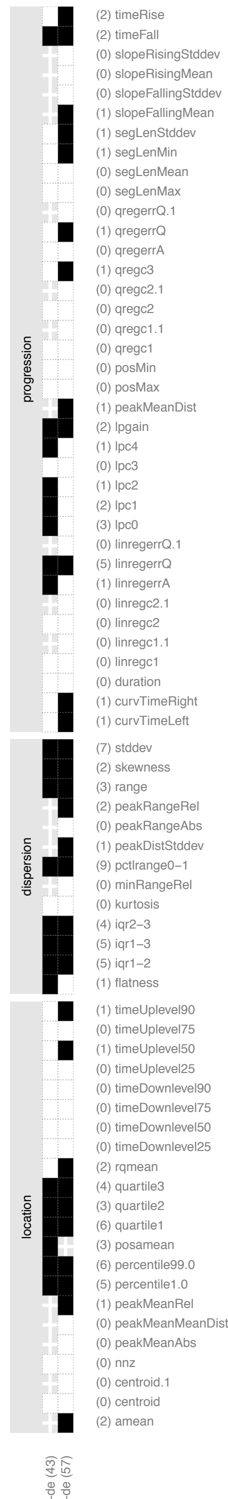
Figure 5.6: Variable selection for personality traits openness (a and b) and conscientiousness (c and d). LLD and Fcn stand for low-level descriptor and functional, respectively. (The number in front of each label indicates number of selected features in that category.)

(a) LLD (b) Fcn (c) LLD (d) Fcn

Figure 5.7: Variable selection for personality traits extraversion (a and b) and agreeableness (c and d). LLD and Fcn stand for low-level descriptor and functional, respectively. (The number in front of each label indicates number of selected features in that category.)



(a) LLD



(b) Fcn

Figure 5.8: Variable selection for personality trait neuroticism. LLD and Fcn stand for low-level descriptor and functional, respectively. (The number in front of each label indicates number of selected features in that category.)

has resulted in a prediction accuracy of 62.5%, which is higher by 2.1% than the case where we use all the 5516 baseline features.

Selected sets of LLDs and functionals for conscientiousness are shown in Figures 5.6c and 5.6d, respectively. According to Figure 5.6c, energy related LLDs are relatively most frequently selected. Next are the spectral LLDs, among which Rasta-filtered spectrum, spectral kurtosis and spectral entropy, and low-end MFCC coefficients have shown to be more relevant to conscientiousness that the rest. According to Figure 5.6d, static functionals are by far more frequently selected than their dynamic counterparts. Among static functionals, location statistics show to be relatively more relevant than spread statistics. Most frequently selected of location statistics selected for modeling conscientiousness are first, second, and third quartiles. Among spread statistics, the three inter-quartile ranges are by far more selected than the rest. Moreover, flatness and standard deviation seem to be informative as well. Among dynamic functional, on the other hand, slope measures and linear prediction gain show to be more relevant than other functionals in the same category.

For the conscientiousness dimension, the reduced set of features comprises 54 LLDs (83%) and 37 functionals (51%), which adds up to 256 features. This is almost 5% of baseline features. This has resulted in a prediction accuracy of 75.5%, which is higher by 1% than the case where we use all the 5516 baseline features.

Selected sets of LLDs and functionals for extraversion are shown in Figures 5.7a and 5.7b, respectively. According to Figure 5.7a, spectral and energy related LLDs seem to be relatively more frequently selected than voicing related ones. Among spectral LLDs, higher order MFCC coefficients and higher bands of Rasta-filtered spectrum show to be more informative than their peers in the same category. According to Figure 5.7b, static functionals show to be more relevant for modeling extraversion than dynamic functionals. Among static functionals spread statistics are relatively more frequently chosen than location statistics. Among spread statistics, inter-quartile ranges, flatness, and kurtosis show to be more relevant than the rest. Among location statistics, on the other hand, quartiles are more frequently chosen than other functionals from the same category. Among dynamic functionals, slope and peaks measures, as well as linear prediction coefficients are among the more relevant functionals from the same category.

For the extraversion dimension, the reduced set of features comprises 36 LLDs (55%) and 25 functionals (35%), which adds up to 70 features. This is less than 2% of baseline features. This has resulted in a prediction accuracy of 86.9%, which is higher by 6% than the case where we use all the 5516 baseline features.

Selected sets of LLDs and functionals for agreeableness are shown in Figures 5.7c and 5.7d, respectively. According to Figure 5.7c, voicing related LLDs show to be more relevant than the other two types for modeling agreeableness. Among LLDs in this category, log harmonic-to-noise ratio, fundamental frequency, shimmer, and jitter show to be more relevant. Among spectral LLDs, MFCC coefficients, frequency bands and lower band Rasta-filtered spectrum, as well as spectral roll-off, skewness, variance, and slope have shown to be relevant to modeling agreeableness. According to Figure 5.7d, dynamic functionals are more frequently selected than their static counterparts. Among dynamic functionals, linear prediction coefficients, positions of maximum and minimum, and rising and falling times show to be more relevant than the rest. Among static functionals, spread statistics are selected relatively more frequently than location statistics. Among spread statistics, inter-quartile range, range, skewness, and standard deviation are more frequently selected. Among location statistics, root quadratic mean, as well as quartiles and percentiles show to be more informative than other types of location statistics.

For the agreeableness dimension, the reduced set of features comprises 59 LLDs (91%) and 57 functionals (79%), which adds up to 340 features. This is almost 6% of baseline features. This has resulted in a prediction accuracy of 66.8%, which is lower by 0.8% than the case where we use all the 5516 baseline features.

Selected sets of LLDs and functionals for neuroticism are shown in Figures 5.8a and 5.8b, respectively. According to Figure 5.8a, voicing related LLDs are relatively more frequently selected than spectral and energy related LLDs. Among voicing related LLDs, fundamental frequency, log harmonic-to-noise ratio, show to be more relevant to neuroticism than the rest. Next are spectral LLDs. Among those, MFCC coefficients, harmonicity, lower end of the Rasta-filtered spectrum, ans spectral flux are chosen more than the rest. According to Figure 5.8b, static functionals are more descriptive for modeling neuroticism. Among those, spread statistics are more frequently selected, among which percentile range, standard deviation, and inter-quartile ranges show to be more relevant. Among location statistics, on the other hand, quartiles and percentiles are selected more frequently than the rest. Among dynamic functionals, linear regression error shows to be most relevant to modeling this dimension of personality trait.

For the neuroticism dimension, the reduced set of features comprises 26 LLDs (40%) and 38 functionals (53%), which adds up to 100 features. This is almost 2% of baseline features. This has resulted in a prediction accuracy of 68.9%, which is higher by 0.9% than the case where we use all the 5516 baseline features.

Discussion

Result of variable selection for 11 different paralinguistic speech recognition tasks are presented. We showed that for most of the studied cases with less than 5% of the features in the original feature vector, we could achieve higher prediction accuracy.

An important detail about the performed analysis in this study is that, although for each selection task some LLDs and functionals are selected, and some others are left out, we may not came to the conclusion that the selected ones are the ones that are useful for the given task, and that the rest are not, given that among two perfectly co-linear features, only one would be useful in a model, and upon choosing one of the two, the other one may be left out. In other words, in the list of discarded LLDs and functionals, there could be some which are as good as the selected ones, however, they are left out in the favor of simplicity, i.e., parsimony. Therefore, the suggested list of features are one of many feasible best choices, according to the criteria under study.

5.2 Max-Dependence Regression

Various regression models are used to predict continuous emotional contents of social signals. The common approach to train those models is minimize a sense of prediction error. According to those optimization criteria, among two models, the one that results in a lower prediction error should be favored. However, prediction error may fall at a lower degree of importance. Instead, linear dependence is commonly used in the literature of affective computing as the measure of goodness of fit. Hence, given that a lower prediction error does not imply a higher dependence, we propose to set maximization of dependence as the optimization criterion. To do so, we take advantage of Hilbert-Schmidt independence criterion as a generic measure of independence. Prediction accuracy of the proposed learning algorithm is compared with that of the support vector regression in the framework of the second audio/visual emotion challenge, as well as by the use of the VAM dataset, and two synthetic datasets.

5.2.1 Background

Regression methods play a pivotal role in the analysis of continuous affect, and a variety of methods are used for this purpose. In principle, regression is an optimization problem that is used to set model parameters, so that the resulting model minimize the prediction error. However, if the criterion for the goodness of a regression model is other than the prediction error, we might accordingly need to modify the optimization criterion for setting the regression coefficients. Particularly, a commonly used measure for assessing the goodness of a prediction in the literature of continuous analysis of affect is correlation coefficient, or analogously linear dependence. In case a model can achieve perfect prediction, that is zero prediction error, that model also does maximizes dependence, however, otherwise, among two models, the one with a lower prediction error does not guarantee a higher dependence. On the other hand, although various models are proved to asymptotically converge to their underlying distributions, achieving zero prediction error in real-world problems, where the amount of observations are far less than asymptote, may not be a realistic target to set. Therefore, when the linear dependence is used as the measure of evaluating a regression model, it would be a fair decision to consequently adapt the optimization problem.

Therefore, we propose a novel regression approach that makes predictions based on a mapping of explanatory variables that maximizes statistical dependencies with the response variable. The maximization identifies a hypersurface along which minimizing the prediction error preserves the maximum dependencies between the mapped explanatory variables and the response variable, resulting in a prediction that is maximally dependant on the response variable. This is in contrast to conventional linear regression approaches, where prediction error is minimized. The conventional approach does not guarantee maximum dependence of the predictions on response variables.

In particular, we distinguish between linear and nonlinear dependencies by using the Hilbert-Schmidt independence criterion (HSIC), a generic measure of dependence, and propose a solution for the regression problem in two stages: 1) extract a set of orthogonal transformations of explanatory variables that maximizes the nonlinear dependency with the response variable, and 2) construct a linear transformation over the mapped explanatory variables that maximizes the linear dependence between these variables and the response variable. HSIC has been previously used for dimensionality reduction [13, 238].

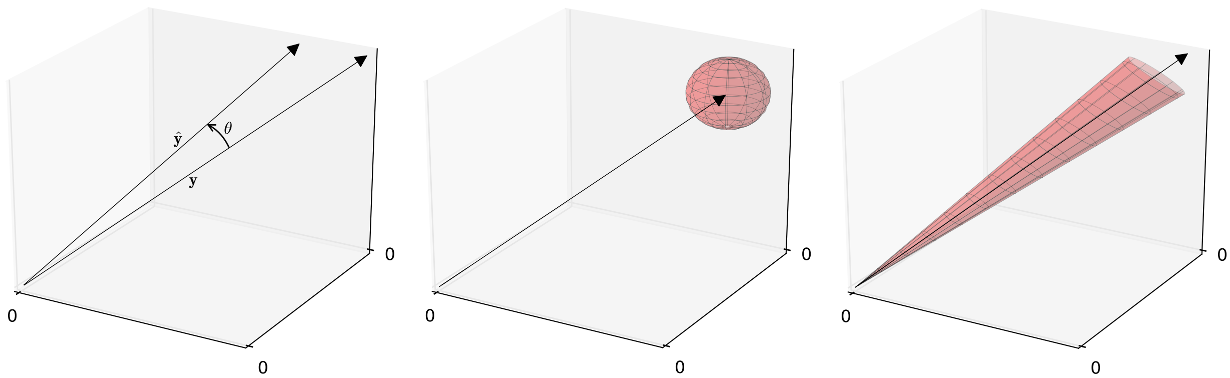
The performance of the proposed approach is evaluated and compared with the state-of-the-art SVR using synthetic datasets. To validate the efficacy of the proposed approach for real-life applications, we apply it to predict affective dimensions for affective speech datasets (AVEC'12 [188] and VAM [92]), and compare the results with those of SVR. Furthermore, synthetic datasets have enabled examining the regression performance at different levels of nonlinearity, noise, and sample size.

5.2.2 Motivation

For two variables $\mathbf{y}$ and $\hat{\mathbf{y}}$ Pearson's correlation coefficient r is defined as

$$r(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sigma_{\mathbf{y}\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}}, \tag{5.2}$$

where σ_y represents the standard deviation of the variable y, and $\sigma_{y\hat{y}}$ denotes the covariance of the two variables y and $\hat{y}$. In the context of vector geometry, r is the cosine of the angle between



(a) (b) (c)

Figure 5.9: (a) Vector geometric interpretation of regression problem (b) Locus of error-minimizing methods (c) locus of correlation-maximizing methods

$\mathbf{y}$ and $\hat{\mathbf{y}}$, θ, where the two vectors intersect at the origin (Figure 5.9a). Therefore, maximizing the correlation coefficient is equivalent to minimizing the angle θ. In this context, what error-minimizing regression algorithms try to achieve is to fix the tip of the vector $\hat{\mathbf{y}}$ in the locus of the points that are of the same distance from the tip of the vector $\mathbf{y}$, i.e., a sphere centered at the end of the vector $\mathbf{y}$ as shown in figure 5.9b, where that distance is proportional to the minimum error. However, this set of points leave loose the angle between the two vectors, that is it could take values in $[0, \arctan \frac{r}{|\mathbf{y}|}]$. Therefore, by decreasing the minimum error, we could not say that correlation would increase. In other words, a sphere with smaller radius does not guarantee smaller angle between the two vectors. And consequently, between two models, the one with the smaller prediction error does not necessarily imply higher correlation. Hence, since projection that minimizes the angle θ is of interest, what we come to propose is to look for the direction of $\hat{\mathbf{y}}$ in the locus of the points that make that angle with the vector $\mathbf{y}$. Those points make for a cone with the aperture of $2\theta^*$ around the vector $\mathbf{y}$, as shown in the Figure 5.9c.

5.2.3 Methodology

Given a set of explanatory variables $\mathbf{x} \in \mathcal{X}$ ($\mathcal{X} \subset \mathbb{R}^p$) and a response variable $y \in \mathcal{Y}$ ($\mathcal{Y} \subset \mathbb{R}$), the objective is to find a dependence-maximizing linear mapping of $\mathcal{X}$ onto $\mathcal{Y}$. This can be formulated as the following optimization problem:

$$\underset{\beta}{\operatorname{argmax}} \quad \text{Dependence}(\mathbf{y}, \mathbf{X}\beta) \tag{5.3}$$

where $\mathbf{y}$ is an $N \times 1$ vector, $\mathbf{X}$ an $N \times p$ matrix, and β a $p \times 1$ vector, with N and p being the number of instances and the number of explanatory variables, respectively. We assume that the explanatory and response variables are standardized, i.e., each variable is normally distributed with a zero mean and standard deviation of one.

First we solve for the maximum correlation solution, that is linear dependence, and then we extend to the general notion of dependence using the Hilbert-Schmidt independence criterion (HSIC). There we get a series of vectors that are highly dependent on the response variable and are linearly independent among themselves. Therefore, to obtain the max-dependence solution, we use the solution obtained by maximizing correlation. In the following, we use lower and uppercase letters to denote scalars, lowercase bold-face to denote vectors, and uppercase bold-face to denote matrices. Moreover, we follow the convention of using Greek letters for parameters, and Latin letters for data.

Pearson Correlation Coefficient

We start by considering the linear dependence criterion, i.e., the Pearson's correlation coefficient.

$$r(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sigma_{\mathbf{y}\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}}, \tag{5.4}$$

where $\sigma_{\mathbf{y}}$ represents the standard deviation of the variable $\mathbf{y}$, and $\sigma_{\mathbf{y}\hat{\mathbf{y}}}$ denotes the covariance of the two variables $\mathbf{y}$ and $\hat{\mathbf{y}}$. Given that we are seeking the linear mapping β that maximizes $r(\mathbf{y}, \hat{\mathbf{y}} = \mathbf{X}\beta)$, we can formulate the optimization problem as follows:

$$\underset{\beta}{\operatorname{argmax}} \quad \frac{\sigma_{\mathbf{y}\mathbf{X}\beta}}{\sigma_{\mathbf{y}}\sigma_{\mathbf{X}\beta}}. \tag{5.5}$$

We can disregard the first term in the denominator, i.e., $\sigma_{\mathbf{y}}$, since it is independent of β. We force the standard deviation of the other term in the denominator to be one, since it only affects the optimal β by a scaling factor. We then have:

$$\begin{aligned} \underset{\beta}{\operatorname{argmax}} \quad & \sigma_{\mathbf{y}\mathbf{X}\beta}, \\ \text{subject to} \quad & \sigma_{\mathbf{X}\beta} = 1. \end{aligned} \tag{5.6}$$

Using Lagrange multipliers and replacing the covariance and standard deviation with their estimates, we have

$$\frac{1}{N-1}\mathbf{y}^\top \mathbf{X}\beta + \lambda(1 - \frac{1}{N-1}\beta^\top \mathbf{X}^\top \mathbf{X}\beta) = 0. \tag{5.7}$$

Then, by taking the derivative with respect to the control parameter β, we have

$$\mathbf{y}^\top\mathbf{X} - 2\lambda\beta^\top\mathbf{X}^\top\mathbf{X} = 0, \tag{5.8}$$

which through some algebraic manipulation leads us to the solution of the optimization problem:

$$\beta_{\text{CC}} \propto (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{5.9}$$

This solution is identical to the solution of the ordinary least squares (OLS) estimator. That is to say, the OLS estimate maximizes the Pearson's correlation coefficient, which could be advantageous due to the well-behaved properties of the OLS, and moreover the variety of methodologies that are developed around ordinary least squares [81].

Despite the upsides of OLS, it is unable to account for a more general sense of dependence. However, if one could capture those dependencies in the form of a number of linearly independent components, then OLS built on those components would be a valid solution to the problem. To address shortcoming of OLS, we consider another notion of independence, the Hilbert-Schmidt independence criterion (HSIC). The promise of HSIC is that it defines dependence in the general sense, since it is established on the kernel spaces of the explanatory and response variables.

Hilbert Schmidt Independence Criterion

Assuming $\mathcal{F}$ and $\mathcal{G}$ to be two separable reproducing kernel Hilbert spaces [83] and $\mathcal{Z} := \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)\} \subseteq \mathcal{X} \times \mathcal{Y}$, HSIC is defined as follows:

$$\begin{aligned} \text{HSIC}(p_{x,y}, \mathcal{F}, \mathcal{G}) = \mathbf{E}_{x,x',y,y'}[k(x,x')l(y,y')] \\ +\mathbf{E}_{x,x'}[k(x,x')]\mathbf{E}_{y,y'}[l(y,y')] \\ -2\mathbf{E}_{x,y}[\mathbf{E}_{x'}[k(x,x')]E_{y'}[l(y,y')]], \end{aligned} \tag{5.10}$$

where pairs of (x, y) are drawn from the joint probability distribution of $\mathcal{X}$ and $\mathcal{Y}$ represented by $p_{x,y}$. $\mathbf{E}$ denotes the expectation operator. To enable approximation given a finite number of samples, the empirical estimate of HSIC [73] is introduced as follows:

$$\text{HSIC}(\mathcal{Z}, \mathcal{F}, \mathcal{G}) = (N-1)^{-2}\mathbf{tr}(\mathbf{KHLH}). \tag{5.11}$$

Where $\mathbf{K}$,$\mathbf{L}$,$\mathbf{H} \in \mathbb{R}^{N\times N}$, $K_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$, $L_{ij} := l(\mathbf{y}_i, \mathbf{y}_j)$, and $H := \mathbf{I} - N^{-1}\mathbf{e}\mathbf{e}^T$, where $\mathbf{e}$ is a vector of N ones. It can be shown that the HSIC of two independent variables is zero. Therefore, by assuming that $\mathbf{K}$ represents a kernel of the linear projection, that is $\mathbf{X}\beta$, and $\mathbf{L}$ a kernel of the response variable $\mathbf{y}$, what is of interest is the mapping β that maximizes $\mathbf{tr}(\mathbf{KHLH})$ [13]. By further assuming that the two kernels are linear, i.e., $\mathbf{K} = \mathbf{X}\beta\beta^\top\mathbf{X}^\top$ and $\mathbf{L} = \mathbf{y}\mathbf{y}^\top$, we have:

$$\begin{aligned}\mathbf{tr}(\mathbf{KHLH}) &= \mathbf{tr}(\mathbf{X}\beta\beta^\top\mathbf{X}^\top\mathbf{H}\mathbf{y}\mathbf{y}^\top\mathbf{H}) \\ &= \mathbf{tr}(\beta^\top\mathbf{X}^\top\mathbf{H}\mathbf{y}\mathbf{y}^\top\mathbf{H}\mathbf{X}\beta)\end{aligned}$$

Hence, we are interested in the solution of the following optimization problem.

$$\begin{aligned}&\underset{\beta}{\text{argmax}} && \mathbf{tr}(\beta^\top\mathbf{Q}\beta), \\ &\text{subject to} && \beta^\top\beta = \mathbf{I},\end{aligned} \tag{5.12}$$

where $\mathbf{Q} = \mathbf{X}^\top\mathbf{HLHX}$. The constraint is to make the optimization problem well defined, since in its absence, it is unbound. Through a set of algebraic manipulations, it can be shown that the solution to this optimization problem is the eigenvectors of $\mathbf{X}^\top\mathbf{HLHX}$ that correspond to the top eigenvalues.

If the kernels are linear, maximizing HSIC is equivalent to maximizing the Pearson's correlation coefficient. Extension to nonlinear kernels is straightforward [13].

Max-Dependence Regression

With the objective of maximizing the dependence between the response variable and the linear mapping of the explanatory variables, as a solution to the regression problem, we propose the following algorithm:

Nonlinear dependence maximization

1. $\mathbf{Q} \leftarrow \mathbf{X}^\top\mathbf{HLHX}$
2. Let columns of $\mathbf{V}$ be the eigenvectors of $\mathbf{Q}$
3. $\beta_{\text{HSIC}} \leftarrow \mathbf{V}_{\mathcal{S}}$, where $\mathcal{S}$ represents the selected subset of $\mathbf{V}$ columns.

Linear dependence maximization

4. $\hat{\mathbf{X}} \leftarrow \mathbf{X}\beta_{\text{HSIC}}$
5. $\beta_{\text{CC}} \leftarrow (\hat{\mathbf{X}}^\top\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^\top\mathbf{y}$

Aggregation

6. $\beta_{\text{MDR}} \leftarrow \beta_{\text{HSIC}} \cdot \beta_{\text{CC}}$

Steps 1-3 encapsulate the required operations for extracting components that are maximally dependent on the response variable. At this stage nonlinear dependence is of interest. Steps 4-5 find a linear combination of the components from the previous stage that maximizes the overall correlation with the response variable. Finally, Step 6 aggregates the two stages.

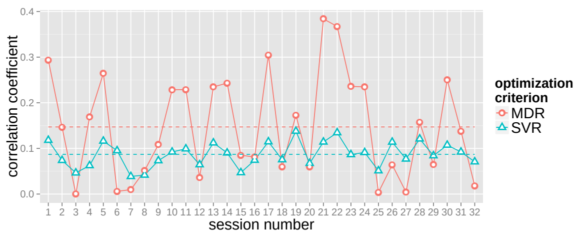
5.2.4 Experimental Study: Induced Affect

Setup

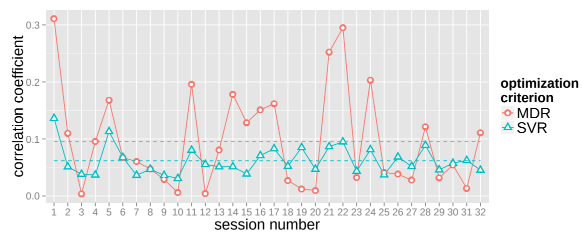
This experiment is designed in the framework of the continuous audio/visual emotion challenge (AVEC) 2012 [188]. Set up in two different time granularities, the challenge targets fully continuous and word-level emotion recognition. In the fully continuous setting, the emotional states of the speakers in every 50 mSec window are estimated, whereas in the word-level setting the emotional content of the expression of each word is of interest. As for the emotional primitives, arousal, expectation, power, and valence are considered. Therefore, the objective of this experiment is to predict continuous affective dimensions from speech samples in the two time granularity. Correlation coefficient (CC), mean absolute percentage error (MAPE), and training and recall times are used to compare the proposed learning algorithm with support vector regression (SVR).

For this experiment, we have used the part of the SEMAINE corpus [123, 124] that was used for the AVEC 2012 [188]. SEMAINE is a database recorded based on the sensitive artificial listener (SAL) interaction scenario [40]. Selected from the recordings of the SEMAINE, AVEC 2012 provides three sets of data, labeled as training, development, and testing sets. Since the response variables for the testing set is not made available to the public, we use the training and development sets for training and hypothesis testing purposes. The number of sessions in the training and development set are 31 and 32, respectively.

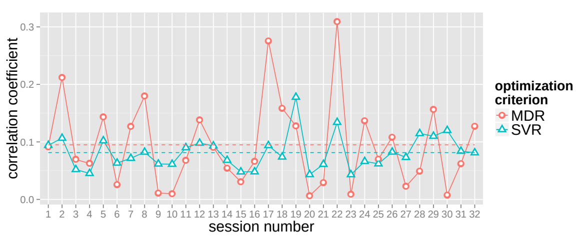
Features are extracted from 100 mSec-long windows of speech signal. The length of the spectral intervals is set to 100 Hz, where two consecutive intervals do not intersect, and collectively they cover 0 to 8 kHz of the spectrum. To set each of these parameters, which are the window size in time domain and the spectral bandwidth, a line search is performed. As for statistics, we use the min, max, median, mean, and standard deviation of the features over windows of a speech signal. This makes a feature vector of 400 dimensions.



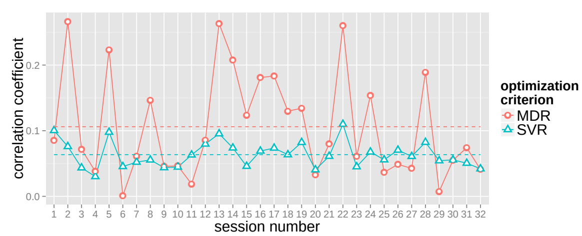
(a) Fully Continuous Emotion Recognition – Arousal



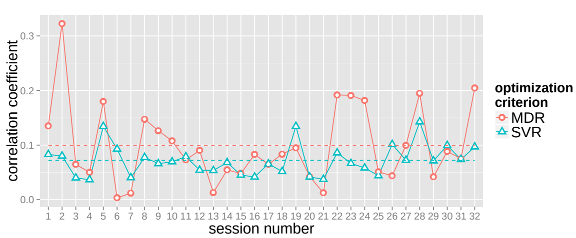
(b) Arousal



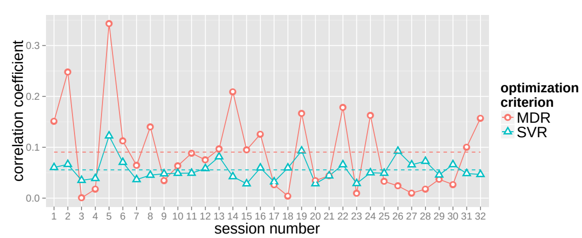
(c) Expectation



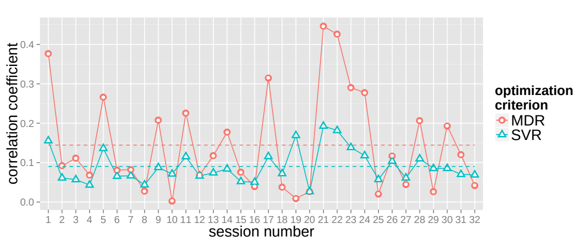
(d) Expectation



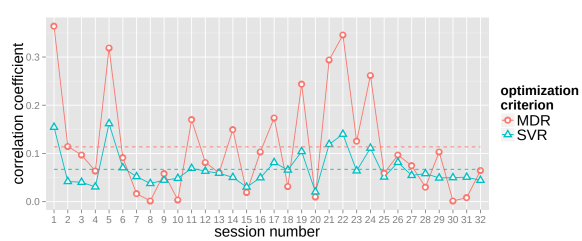
(e) Power



(f) Power

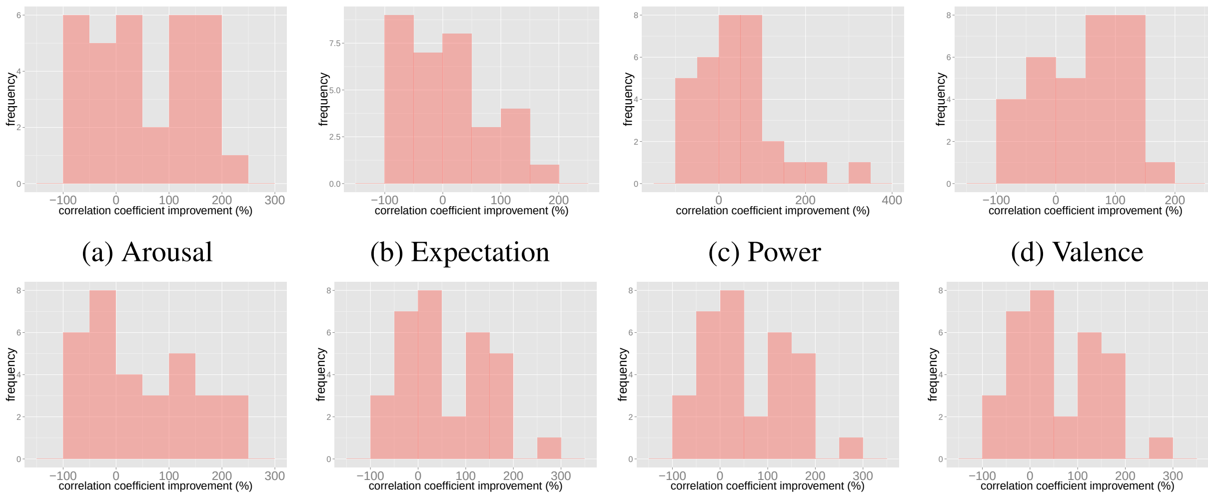


(g) Valence



(h) Valence

Figure 5.10: Correlation coefficient of the predictions with the actual values, per session of the development set for fully continuous (first column) and word-level (second column) recognition of affect. The dashed lines indicate the average correlation coefficient of predicted values with the actual response values, for each method and over all the sessions.



(a) Arousal (b) Expectation (c) Power (d) Valence

(e) Arousal (f) Expectation (g) Power (h) Valence

Figure 5.11: Distribution of the development sessions population in terms of the relative correlation coefficient advantage of MDR over SVR, for fully continuous (first row) and word-level (second row) recognition of affect.

Results and Discussion

We have used each of the models to predict the emotional content of each of the 32 sessions of the development set. Correlation coefficient of the predictions with the actual emotional content of each of the sessions is shown in Figure 5.10. The dashed lines in these figures show the average correlation coefficient of the prediction over all the sessions. For the arousal dimension, for both time granularities, we notice that MDR results in higher average correlation coefficients than SVR. Similarly, for the expectation dimension, MDR outperforms SVR with respect to the correlation coefficient of their prediction, however in the case of the fully continuous recognition, the differences is not as noticeable. Moreover, for the power and valence emotion dimensions, for both time granularities, MDR results in higher correlation coefficient than SVR.

Looking closer at Figure 5.10, we notice that for most of the sessions MDR performs better than SVR, however, there are some sessions and affective dimensions for which SVR results in higher correlation coefficient. To show these differences more clearly, we use the distribution of the relative advantage of MDR with respect to SVR. What we mean by relative advantage is the correlation difference between MDR and SVR, normalized by the correlation of the SVR. Those distributions are shown in Figure 5.11. In Table 5.3, the average relative advantage of MDR with respect to SVR is recorded. According to this information, MDR outperforms SVR in all

Table 5.3: The relative advantage of MDR with respect to SVR.

Arousal	Expectation	Power	Valence
Fully Continuous Emotion Recognition			
69%	17%	38%	60%
Word-Level Emotion Recognition			
56%	67%	63%	69%

Table 5.4: Comparison of the performance of MDR and SVR, in terms of the correlation coefficient (CC) and the mean absolute percentage error (MAPE) of the predicted values, as well as the training and recall time (T_T and T_R) in millisecond.

	Arousal		Expectation		Power		Valence		Average			
Method	CC	MAPE	CC	MAPE	CC	MAPE	CC	MAPE	CC	MAPE	T_T	T_R
Fully Continuous Emotion Recognition												
MDR	**0.147**	48.78	**0.095**	211.75	**0.099**	25.85	**0.144**	70.13	**0.121**	89.13	**1416**	**1**
SVR	0.087	50.55	0.081	215.13	0.072	26.78	0.090	70.80	0.082	90.82	23968	60
Word-Level Emotion Recognition												
MDR	**0.096**	48.30	**0.106**	70.07	**0.091**	24.97	**0.114**	64.86	**0.102**	52.05	**3304**	**2**
SVR	0.062	49.76	0.064	70.25	0.056	25.82	0.067	65.41	0.062	52.81	74893	174

eight cases. Furthermore, except the expectation and power dimensions in the fully continuous recognition task, for the other six cases, which are arousal and valence for both time granularities and expectation and power for word-level recognition, MDR is more than 50% advantageous to the SVR.

A valid question here is how MDR perform in terms of the prediction error. To answer to this question, we have included the average mean absolute percentage error (MAPE) of the prediction over all the sessions, together with the average correlation coefficients of the predictions in Table 5.4. There, one can notice that compared to SVR, MDR results in higher correlation coefficient and lower prediction error. Moreover, in the same table the average training and recall times (T_T and T_R) are included. According to these numbers, MDR is, by an order of 4 to 23 time for training and 60 to 87 times for recall, faster than SVR.

5.2.5 Experimental Study: Spontaneous Affect

Setup

The performance of the proposed max-dependence regression (MDR) approach is evaluated and compared with that of support vector regression (SVR) using 10 repetitions of 10-fold cross validation (referred to as 10×10FCV, hereafter) on the VAM dataset. 10-fold cross validation is used for its reliability in model selection and accuracy estimation [103, 165]. Moreover, same settings for the folds are used to test the performances of MDR and SVR in the 10×10FCV. Correlation coefficient (CC) and mean absolute percentage error (MAPE) are used for evaluation. Additionally, training and recall times are used to compare the computational complexity of the algorithms.

There are different sources of variation in expression of affect, among which are person-specific and idiosyncratic variations. In order to test the generalization ability of the proposed approach to different subjects, leave-one-subject-out cross validation (LOSOCV) is also performed. In each fold of LOSOCV, a subject is left out (testing subject) and the models are trained using the remaining subjects (training subjects).

In these experiments, the kernelized versions of MDR and SVR are used, where we have considered linear, radial basis, and polynomial kernels. The kernel types and their hyper-parameters for MDR and SVR and the SVR's slack parameter are selected to optimize Pearson's correlation coefficient in a cross validation test performed on the training set[1].

Results and Discussion

Table 5.5 shows average CC(± std) and MAPE(± std) for the predicted affective dimensions obtained by 10×10FCV. For both MDR and SVR, the best results were obtained by the radial basis kernel. The high CC and low MAPE resulting from both approaches in the prediction of the activation and dominance dimensions show high accordance of the predicted values with those perceived by the observers.

Unlike the activation and dominance dimensions, for which MDR and SVR perform equally well, the performance is significantly poorer for valence. This is in spite of the fact that the Cronbach's alpha for valence is in the *good* range, meaning that the agreement between the observers is relatively high. A possible explanation is that the observers' evaluation is based on both the audio and visual modalities, and that the two modalities are not equally effective

[1]In each fold of 10×10FCV and LOSO, a separate 5-fold cross validation is performed using only the training set, and kernels (and their hyper-parameters) maximizing CC are selected to perform regression in that fold.

Table 5.5: Results on the affective speech dataset. (CC: Correlation Coefficient, MAPE: Mean Absolute Percentage Error)

	Activation		Dominance		Valence	
	CC	MAPE	CC	MAPE	CC	MAPE
			10×10-Fold Cross Validation			
SVR	82.08 ± 0.45	5.92 ± 0.04	76.10 ± 0.45	5.64 ± 0.05	46.72 ± 1.29	8.04 ± 0.14
MDR	82.15 ± 0.29	6.01 ± 0.05	77.36 ± 0.26	5.72 ± 0.03	43.43 ± 1.62	9.30 ± 0.11
			Leave-One-Subject-Out			
SVR	81.68	6.00	74.95	5.76	40.83	8.59
MDR	81.23	6.17	75.07	5.99	33.09	9.79

in conveying different dimensions of affect. Since only the audio part of the dataset is used for regression, the low level of correlation in predicting valence might be due to the insufficiency of the explanatory variables.

To further examine the capability of the proposed approach in generalizing to unseen subjects, leave-one-subject-out cross validation experiments are conducted. The results of those experiments are shown in terms of the CC and MAPE in Table 5.5. The trend here is very similar to that of the 10×10FCV, where both MDR and SVR show similar performance in predicting the activation and dominance dimensions of unseen subjects, and considerably lower performance in predicting the valence. Although the two approaches do not show a meaningful difference in predicting activation and dominance, the difference is noticeable for the valence.

5.2.6 Experimental Study: Synthetic Datasets

Setup

The performance of the proposed max-dependence regression (MDR) approach is evaluated and compared with that of support vector regression (SVR) using 10 repetitions of 10-fold cross validation (referred to as 10×10FCV, hereafter) on synthetic datasets. 10-fold cross validation is used for its reliability in model selection and accuracy estimation [103, 165]. Moreover, same settings for the folds are used to test the performances of MDR and SVR in the 10×10FCV.

In these experiments, the kernelized versions of MDR and SVR are used, where we have considered linear, radial basis, and polynomial kernels. The kernel types and their hyper-parameters for MDR and SVR and the SVR's slack parameter are selected to optimize Pearson's correlation coefficient in a cross validation test performed on the training set[2].

The performance of the proposed max-dependence regression (MDR) approach is evaluated and compared with that of support vector regression (SVR) using 10 repetitions of 10-fold cross validation (referred to as 10×10FCV, hereafter) on the synthetic and real datasets. 10-fold cross validation is used for its reliability in model selection and accuracy estimation [103, 165]. Moreover, same settings for the folds are used to test the performances of MDR and SVR in the 10×10FCV. Cross correlation (CC) and mean absolute percentage error (MAPE) are used for evaluation. Additionally, training and recall times are used to compare the computational complexity of the algorithms.

Synthetic datasets were used to enable assessment of the proposed approach at varying levels of non-linearity, noise, and size. Different synthetic datasets were tested. For the sake of brevity, we present results from two of these datasets:

1. **Dataset.1:** the regression model is defined as:
$$Y = X_1(\sin(2\pi f X_2) + 1) + \epsilon, \tag{5.13}$$
2. **Dataset.2:** the regression model is defined as:
$$Y = \mathrm{sinc}(2\pi f X_2) + \epsilon. \tag{5.14}$$

In these synthetic datasets, $X_1 \in [0, 1]$ and $X_2 \in [0, 1]$ are uniformly distributed random variables, f is the frequency, and ϵ is a normal additive noise $\epsilon \sim N(0, \sigma^2)$. σ^2 is the variance of the normal noise.

For these datasets, data are generated using different combinations of 3 sample sizes (50, 100, 500), 10 frequencies (0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 128, 1024), and 5 noise ratio levels (0.0125, 0.025, 0.05, 0.5, 1), for a total of 150 cases. Additionally, to equalize conditions under which MDR and SVR are compared, the linear implementation of both approaches is tested. It is clear that a kernelized implementation of MDR and SVR is more suited for the cases where a high-level of nonlinearity is introduced in the synthetic datasets. However, implementing kernelized MDR and SVR requires selecting a suitable kernel and tuning its hyper-parameters, which adds to the number of conditions under which the two approaches are compared. While we recognize the importance of comparing the relative performance of the kernelized MDR and SVR with synthetic nonlinear datasets, for the sake of simplicity of analysis (number of comparison conditions) and to maintain an equal ground for comparing the two approaches, we only implement linear versions of the two approaches in the present work. An extended comparison based on kernelized MDR and SVR with synthetic nonlinear datasets is a future direction for this work.

[2] In each fold of 10×10FCV, a separate 5-fold cross validation is performed using only the training set, and kernels (and their hyper-parameters) maximizing CC are selected to perform regression in that fold.

Results and Discussion

The results on the two synthetic datasets are shown in Figure 5.12, where each point corresponds to an experiment with samples generated given a sample size, a frequency, and a noise ratio, and abscissa and ordinate of each point indicate the resulting correlation coefficient by MDR and SVR, respectively. We use the relative position of the points with respect to the identity (1:1) line to assess the relative performance of the two approaches in each scenario. Points that are on the top side of the line favor MDR over SVR, and points that are on the bottom side favor SVR over MDR. The further a point gets from the line, the more one approach is favored over the other.

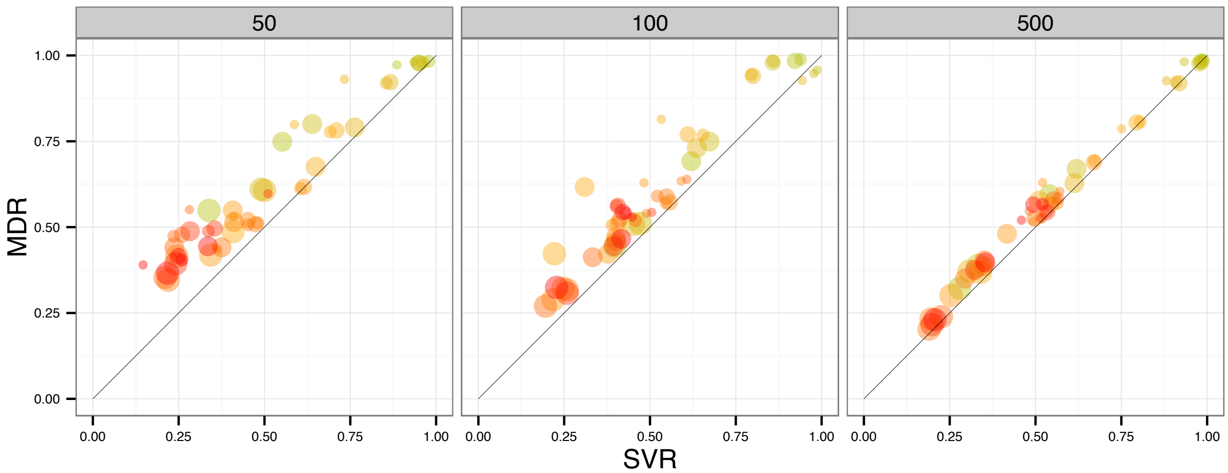
Figure 5.12a presents the results of the first synthetic dataset (Equation 5.13). According to this figure, in more than 95% of cases (143/150), MDR produces higher correlation than SVR. For 50 samples, MDR shows better performance than SVR in all cases. By increasing the sample size, we see an evident shift towards the identity line, and despite the better performance of MDR in the sample size of 500, points are very close to the identity line. The sum of distances of the points to the identity line for the sample sizes of 50, 100, and 500, are 4.03, 3.13, and 1.12. For this dataset, we could say that MDR shows better performance compared to SVR when few data points are available.

Increasing the frequency and/or noise ratio, decreases the overall performance of both methods. This is expected, given that these two parameters contribute to nonlinearity and unpredictability of data, respectively. However, the degree to which the two methods are affected by these changes is different. Figure 5.13 shows the relative trend of changes of correlation with respect to sample size, frequency, and noise ratio. The ordinate of these figures indicates the percentage of cases where MDR results in a higher correlation than SVR. As the frequency or noise ratio increase, MDR's performance monotonously becomes better than that of SVR.

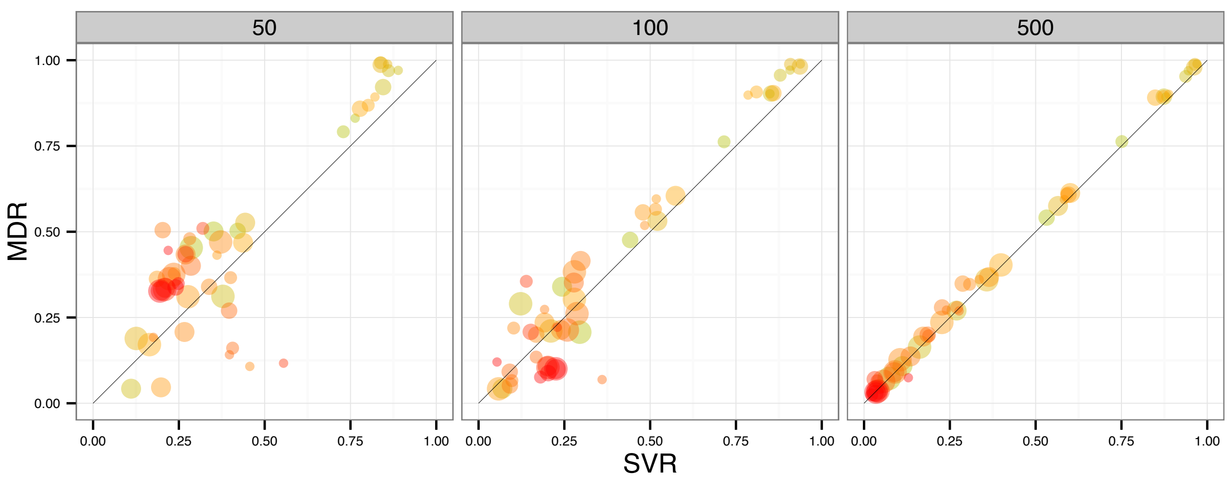
Average training and recall times are 106.4 and 0.6 milliseconds for MDR, and 2700.5 and 0.8 milliseconds for SVR. That is, MDR is more than 25 times faster than SVR in terms of training time, with similar recall time.

Figure 5.12b presents the results of the second synthetic dataset (Equation 5.14). According to this figure, in more than 75% of cases (113/150), MDR results in a higher correlation than SVR. In terms of sample size, a similar trend to the first synthetic dataset is observed. For the lowest sample size, MDR outperforms SVR, with increased sample size, they tend to show more similar results, still in favor of MDR for the higher sample sizes. The sum of distances of the points to the identity line for the samples sizes of 50, 100, and 500, are 4.47, 2.49, and 0.54.

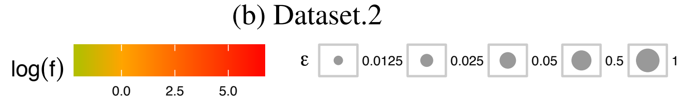
For this dataset, the trend does not seem to be as smooth (Fig. 5.13), however, in this case too the average correlation for MDR is higher than that of SVR. Despite the performance deterioration with the increase in the frequency and noise ratio, MDR results in a higher correlation than



(a) Dataset.1

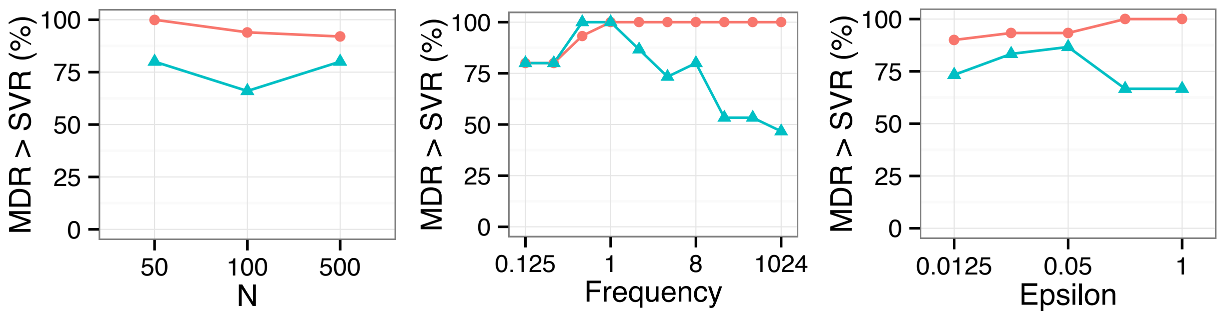


(b) Dataset.2



(c)

Figure 5.12: Relative performance of MDR versus SVR for different combinations of sample size (50, 100, and 500), frequency (f), and noise ratio (ϵ). Points that are above (below) the line are those that favor MDR (SVR).



(a) (b) (c)

Figure 5.13: Trends of changes of CC with respect to sample size, frequency, and noise ratio. Disk and triangle correspond to the dataset.1 and 2, respectively.

that of SVR. The only exception is the results at $f = 1024$, where SVR demonstrates a slightly better performance (Figure 5.13). However, the correlation coefficients at this frequency in all cases fall below 30% for both approaches, hence, no conclusion can be derived on the superiority of one approach over the other.

Average training and recall times are 108.3 and 0.6 milliseconds for MDR, and 1465.1 and 0.8 milliseconds for SVR. That is, SVR is more than 13 times slower than MDR in terms of training time, however, they are similarly fast in the recall phase.

Discussion

The experiments with the synthetic datasets were designed to evaluate the relative performance of MDR and SVR at varying levels of non-linearity, unpredictability, and sample size. Non-linearity and unpredictability were introduced by varying frequency and noise ratio, respectively. Based on the results reported in Section 5.2.6, we can make the following hypotheses: 1) MDR outperforms SVR when few samples are available and the two approaches perform more similarly as the sample size increases, 2) SVR performs better than MDR at smaller noise ratios; at higher noise ratios MDR outperforms SVR, 3) the performance of both approaches deteriorates as the frequency (viz. nonlinearity) increases. The third hypothesis is weak due to the lack of experiments with kernelized MDR and SVR for the nonlinear datasets in the present work.

To further evaluate the first hypothesis, additional experiments were conducted using the VAM dataset. A 10×10 FCV was conducted using 20% of samples randomly selected from VAM Audio-I. The results from the experiment with a subset of VAM Audio-I dataset show that SVR outperforms MDR in terms of average cross-validated CC's (SVR $>$ MDR by: 2.83% in

Activation, 7.06% in Dominance, and 2.51% in Valence). These results do not support the first hypothesis on the advantage of MDR over SVR for small sample sizes.

A similar poorer performance of MDR in comparison with SVR is also observed in the experiments with 50 samples of the synthetic dataset 2. As can been seen in Figure 5.12a, in all such cases, the noise ratio is very low and as the noise ratio increases and sample size remains fixed, MDR surpasses SVR. A possible explanation is that there is an interaction effect between the noise ratio and sample size such that the effect of sample size varies at different levels of noise ratio. To test this hypothesis, we have rerun the 10×10 FCV with 20% of samples randomly selected from VAM Audio-II where there is a lower agreement between observers on conveyed affective dimensions in comparison with VAM Audio-I; which in turn makes it more noisy than VAM Audio-I. VAM Audio-II contains 469 samples in total. On average, MDR performs better than SVR on the subset of VAM Audio-II dataset in terms of cross-validated CC's from 10×10 FCV (MDR > SVR by: 2.19% in Activation, 1.29% in Dominance, 1.66% in Valence).

Therefore, the relative performance of MDR and SVR on subsets of VAM Audio-I and VAM Audio-II shows that at a similar sample size, SVR outperforms MDR at lower noise ratios (VAM Audio-I), while at higher noise ratios (VAM Audio-II), MDR outperforms SVR, which is congruent with the hypothesis on the interaction effect of noise ratio and sample size. These results also support the hypothesis regarding MDR's advantage at higher noise ratios (Hypothesis 2).

Another advantage of MDR over SVR is its computational efficiency. MDR's training and recall times for the synthetic and real datasets are significantly shorter than those of SVR.

Another important observation is that by decreasing the number of explanatory variables from two (synthetic dataset 1) to one (synthetic dataset 2), the average training time of SVR is almost halved (from 2700.5 ms in dataset 1 to 1465.1 ms in dataset 2), whereas MDR's training time did not meaningfully change (106.4 ms in dataset 1 and 108.3 ms in dataset 2). The importance of this difference could be even more evident in cases where the dimensionality of the feature space is large.

5.3 Dictionary Learning

Recently, a supervised dictionary learning approach based on the Hilbert-Schmidt independence criterion has been proposed that learns the dictionary and the corresponding sparse coefficients in a space where the dependency between the data and the corresponding labels is maximized. In this section, two multi-view dictionary learning techniques are proposed based on that supervised dictionary learning approach. While one of these two techniques learns one dictionary and the corresponding coefficients in the space of fused features in all views, the other learns one dictionary in each view and subsequently fuses the sparse coefficients in the spaces of learned dictionaries. The effectiveness of the proposed multi-view learning techniques in using the complementary information of single views is demonstrated in the application of affective speech recognition. The fully-continuous sub-challenge of the audio/visual emotion challenge dataset is used in two different views: baseline and spectral energy distribution feature sets. Four dimensional affects, i.e., arousal, expectation, power, and valence are predicted using the proposed multi-view methods as the continuous response variables. Results are compared with the single-views baseline results of the challenge, and also other supervised and unsupervised multi-view learning approaches in the literature. Using correlation coefficient as the performance measure in predicting the continuous dimensional affects, it is shown that the proposed approach achieves the highest performance among all the considered models.

5.3.1 Background

There are many mathematical models with varying degrees of success to describe data, among which dictionary learning and sparse representation (DLSR) have attracted the interest of many researchers in various fields. Dictionary learning and sparse representation are two closely-related topics that have roots in the decomposition of signals to some *predefined* bases, e.g., Fourier transform. Representation of signals using predefined bases is based on the assumption that these bases are general enough to represent any kind of signal, however, recent research shows that learning bases from data leads to state-of-the-art results in many applications such as texture classification [64, 208, 228], face recognition [221, 231, 240], image denoising [43, 120], biomedical tissue characterization [62, 66, 196], motion and data segmentation [45, 163], data representation and column selection [44], and image super-resolution [230]. What makes DLSR distinct from the representation using predefined bases can be summarized as follows: firstly, bases are learned from data, and secondly, only a few components in the dictionary are needed to represent the data, i.e., sparse representation. This latter attribute can also be seen in the decomposition of signals using some predefined bases such as wavelets [121].

For a more formal description, let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d\times N}$ be a finite set of data samples, where d is the dimensionality and N is the number of data samples. The main goal in classical dictionary learning and sparse representation is to decompose the data over a few dictionary atoms by minimizing a loss function as follows

$$L(\mathbf{X}, \mathbf{D}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} l(\mathbf{x}_i, \mathbf{D}, \boldsymbol{\alpha}), \tag{5.15}$$

where $\mathbf{D} \in \mathbb{R}^{d\times l}$ is the dictionary of l atoms, and $\boldsymbol{\alpha} \in \mathbb{R}^{l\times N}$ are the coefficients. The most common loss function in the DLSR literature is the reconstruction error between the original data samples $\mathbf{X}$ and the decomposed data in the space of the learned dictionary $\mathbf{D}$, regularized using a sparsity inducing function to guarantee the sparsity of the coefficients. The most common sparsity inducing function is ℓ_1 norm. Hence, (5.15) can be rewritten as

$$L(\mathbf{X}, \mathbf{D}, \boldsymbol{\alpha}) = \min_{\mathbf{D}, \boldsymbol{\alpha}} \sum_{i=1}^{N} \left(\tfrac{1}{2} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1 \right), \tag{5.16}$$

where $\boldsymbol{\alpha}_i$ is the i^{th} column of $\boldsymbol{\alpha}$.

The optimization problem in (5.16) is mainly based on the minimization of the reconstruction error in mean-squared sense, which is optimal in applications such as denoising, inpainting, and coding [86]. However, the representation obtained from (5.16) does not necessarily lead to a discriminative representation, which is important in classification tasks.

Several approaches have recently been proposed to include class labels into the optimization problem given in (5.16) to yield a discriminative representations. These approaches can broadly be grouped into five categories as suggested in [65][3], including: 1) Learning one dictionary per class, where one sub-dictionary is learned per class and then all the sub-dictionaries are composed into one. Supervised *k*-means [207, 208], sparse representation-based classification (SRC) [221], metaface [232], and dictionary learning with structured incoherence (DLSI) [162] are in this category. 2) Pruning large dictionaries, in which, initially a very large dictionary is learned in an unsupervised manner, and then the atoms in the dictionary are merged according to some objective function that takes into account the class labels. The supervised dictionary learning approaches based on agglomerative information bottleneck (AIB) [60] and universal visual dictionary [216] are in this category. 3) Learning the dictionary and classifier in one optimization problem, where the optimization problem for the classifier is embedded into the optimization problem given in (5.16) or its modified version. Discriminative SDL [119] and discriminative K-SVD (DK-SVD) [237] are two techniques in this category. 4) Including class labels in the learning of the dictionary, such as the technique based on information loss minimization (known as info-loss) [108] and the one based on randomized clustering forests (RCF) [132]. 5) Including class labels in the learning of the sparse coefficients or both the dictionary and coefficients such as Fisher discrimination dictionary learning (FDDL) [231].

Recently, a supervised dictionary learning approach has been proposed [65] which is based on the Hilbert Schmidt independence criterion (HSIC) [72], in which the category information is incorporated into the dictionary by learning the dictionary in a space where the dependency between the data and class labels is maximized. The approach has several attractive features

[3]The interested reader is encouraged to refer to [65] and the references thereof for a more extensive review on various supervised dictionary learning approaches in the literature and their main advantages and shortcomings.

such as closed-form formulation for both the dictionary and sparse coefficients, very compact dictionary, i.e., discriminative dictionary at small size, and fast efficient algorithm [65]. Thus, it has been adopted in this study.

There are instances where the data in a dataset is represented in multiple views [15]. This can be due to the availability of several feature sets for the same data such as representation of a document in several languages [3], representation of webpages by both their text and hyperlinks, etc., or due to the availability of information from several modalities, e.g., biometric information for the purpose of authentication that may come from fingerprints, iris, and face. Although single-view representation might be sufficient in a machine learning task for the analysis of the data, complementary information provided by multiple views usually facilitates the improvement of the learning process.

In this paper, we provide the formulation for multi-view learning based on the supervised dictionary learning proposed in [65]. Two different methods for multi-view representation are proposed and the application to affective speech recognition using two different feature sets are investigated. Additionally, the multi-view approach is extended to continuous labels, i.e., to the case of a regression problem (it was originally proposed for classification tasks using discrete labels [65]). It is worth to note that not all the proposed supervised dictionary learning approaches in the literature can be extended to regression problems. For example, in supervised *k*-means, the discrete labels are needed and it cannot be extended to continuous labels. We will show that the proposed approach can effectively use the complementary information in different feature sets and improve the performance of the recognition system on the AVEC (audio/visual emotion challenge) 2012 emotion recognition dataset compared with some other supervised and unsupervised multi-view approaches.

5.3.2 Methodology

In this section, the formulation of the proposed multi-view supervised dictionary learning (MV-SDL) is provided. To this end, we first briefly review the Hilbert-Schmidt independence criterion (HSIC). Then we provide the formulation for the adopted supervised dictionary learning as being proposed in [65]. Eventually, the mathematical formulation for the proposed MV-SDL is presented.

Hilbert-Schmidt Independence Criterion

HSIC is a kernel-based independence measure between two random variables $\mathcal{X}$ and $\mathcal{Y}$ [72]. It computes the Hilbert-Schmidt norm of the cross-covariance operators in reproducing kernel

Hilbert Spaces (RKHSs) [6, 72].

Suppose that $\mathcal{H}$ and $\mathcal{G}$ are two RKHSs in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Hence, by the Riesz representation theorem, there are feature mappings $\phi(x) : \mathcal{X} \to \mathbb{R}$ and $\psi(y) : \mathcal{Y} \to \mathbb{R}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ and $l(y, y') = \langle \psi(y), \psi(y') \rangle_{\mathcal{G}}$.

HSIC can be practically estimated in the RKHSs using a finite number of data samples. Let $\mathcal{Z} := \{(\mathbf{x}_1, \mathbf{y}_1,), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\} \subseteq \mathcal{X} \times \mathcal{Y}$ be N independent observations drawn from $p := P_{\mathcal{X} \times \mathcal{Y}}$. The empirical estimate of HSIC can be computed using [72]

$$\text{HSIC}(\mathcal{Z}) = \frac{1}{(N-1)^2} \text{tr}(\mathbf{KHLH}), \tag{5.17}$$

where tr is the trace operator, $\mathbf{H}, \mathbf{K}, \mathbf{L} \in \mathbb{R}^{N \times N}$, $K_{i,j} = k(x_i, x_j)$, $L_{i,j} = l(y_i, y_j)$, and $\mathbf{H} = \mathbf{I} - N^{-1}\mathbf{e}\mathbf{e}^\top$ ($\mathbf{I}$ is the identity matrix, and $\mathbf{e}$ is a vector of N ones, and hence, $\mathbf{H}$ is the centering matrix). According to (5.17), maximizing the empirical estimate of HSIC, i.e., $\text{tr}(\mathbf{KHLH})$, will lead to the maximization of the dependency between two random variables $\mathcal{X}$ and $\mathcal{Y}$.

HSIC-Based Supervised Dictionary Learning

The HSIC-based supervised dictionary learning (SDL) learns the dictionary in a space where the dependency between the data and corresponding class labels is maximized. To this end, it has been proposed in [65] to solve the following optimization problem

$$\begin{aligned} \max_{\mathbf{U}} \quad & \text{tr}(\mathbf{U}^\top \mathbf{XHLHX}^\top \mathbf{U}), \\ \text{s.t.} \quad & \mathbf{U}^\top \mathbf{U} = \mathbf{I} \end{aligned} \tag{5.18}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ is N data samples with the dimensionality of d; $\mathbf{H}$ is the centering matrix; $\mathbf{L}$ is a kernel on the labels $\mathbf{y}$; and $\mathbf{U}$ is the transformation that maps the data to the space of maximum dependency with the labels. According to the Rayleigh-Ritz Theorem [118], the solution for the optimization problem given in (5.18) is the corresponding eigenvectors of the top eigenvalues of $\mathbf{\Phi} = \mathbf{XHLHX}^\top$.

To explain how the optimization problem provided in (5.18) learns the dictionary in the space of maximum dependency with the labels, using a few manipulations, we note that the objective function given in (5.18) has the form of empirical HSIC given in (5.17), i.e.,

$$
\begin{aligned}
\max_{\mathbf{U}} \operatorname{tr}(\mathbf{U}^\top \mathbf{X}\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{X}^\top \mathbf{U}) \\
&= \max_{\mathbf{U}} \operatorname{tr}(\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}\mathbf{H}\mathbf{L}\mathbf{H}) \\
&= \max_{\mathbf{U}} \operatorname{tr}\bigg(\Big[(\mathbf{U}^\top \mathbf{X})^\top \mathbf{U}^\top \mathbf{X}\Big]\mathbf{H}\mathbf{L}\mathbf{H}\bigg) \\
&= \max_{\mathbf{U}} \operatorname{tr}(\mathbf{K}\mathbf{H}\mathbf{L}\mathbf{H}), \qquad (5.19)
\end{aligned}
$$

where $\mathbf{K} = (\mathbf{U}^\top \mathbf{X})^\top \mathbf{U}^\top \mathbf{X}$ is a linear kernel on the transformed data in the subspace $\mathbf{U}^\top \mathbf{X}$. To derive (5.19), it is noted that the trace operator is invariant under cyclic permutation, e.g., $\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA)$ and also that $\mathbf{X}^\top \mathbf{U} = (\mathbf{U}^\top \mathbf{X})^\top$.

Now, it is easy to observe that the form given in (5.19) is the same as empirical HSIC in (5.17) up to a constant factor and therefore, it can be easily interpreted as transforming centered data $\mathbf{X}$ using the transformation $\mathbf{U}$ to a space where the dependency between the data and class labels is maximized. In other words, the computed transformation $\mathbf{U}$ constructs the dictionary learned in the space of maximum dependency between the data and class labels.

After finding the dictionary $\mathbf{D} = \mathbf{U}$, the sparse coefficients can be computed using the formulation given in (5.16). As explained in [65], (5.16) can be either solved using iterative methods such as the *lasso* or in closed-form using soft-thresholding [39, 59] with the soft-thresholding operator $S_\lambda(.)$, that is,

$$\alpha_{ij} = S_\lambda\Big([\mathbf{D}^\top \mathbf{x}_i]_j\Big), \qquad (5.20)$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is the i^{th} data sample, $[\mathbf{D}^\top \mathbf{x}_i]_j$ and α_{ij} are the j^{th} elements of $\mathbf{D}^\top \mathbf{x}_i$ and $\boldsymbol{\alpha}_i$, respectively, and $S_\lambda(t)$ is defined as follows

$$S_\lambda(t) = \begin{cases} t - 0.5\lambda & \text{if } t > 0.5\lambda \\ t + 0.5\lambda & \text{if } t < -0.5\lambda \\ 0 & \text{otherwise} \end{cases}$$

The steps for the computation of the dictionary and coefficients using the HSIC-based SDL is provided in Algorithm 1.

The main advantages of the HSIC-based SDL are that the dictionary and coefficients are computed in closed form and separately. Hence, unlike many other SDL techniques in the literature, learning these two do not have to be performed iteratively and alternately. Another remark on the

Algorithm 1 HSIC-Based Supervised Dictionary Learning [65]

Input: Training data, $\mathbf{X}_{\text{tr}}$, test data, $\mathbf{X}_{\text{ts}}$, kernel matrix of labels $\mathbf{L}$, training data size, N, size of dictionary, l.

Output: Dictionary, $\mathbf{D}$, coefficients for training and test data, $\boldsymbol{\alpha}_{\text{tr}}$ and $\boldsymbol{\alpha}_{\text{ts}}$.

1: $\mathbf{H} \leftarrow \mathbf{I} - N^{-1}\mathbf{e}\mathbf{e}^{\top}$

2: $\boldsymbol{\Phi} \leftarrow \mathbf{X}_{\text{tr}}\mathbf{HLHX}_{\text{tr}}^{\top}$

3: **Compute Dictionary:** $\mathbf{D} \leftarrow$ eigenvectors of $\boldsymbol{\Phi}$ corresponding to top l eigenvalues

4: **Compute Training Coefficients:** For each data sample $\mathbf{x}_{\text{tr}_i}$ in the training set, use $\alpha_{ij} = S_{\lambda}\left([\mathbf{D}^{\top}\mathbf{x}_{\text{tr}_i}]_j\right), j = 1, ..., l$ to compute the corresponding coefficient

5: **Compute Test Coefficients:** For each data sample $\mathbf{x}_{\text{ts}_i}$ in the test set, use $\alpha_{ij} = S_{\lambda}([\mathbf{D}^{\top}\mathbf{x}_{\text{ts}_i}]_j), j = 1, ..., l$ to compute the corresponding coefficient

HSIC-based SDL is that unlike many other SDLs in the literature, the labels $\mathbf{y}$ are not restricted to discrete values and can also be continuous. In other words, the HSIC-based SDL can be easily extended to regression problems, in which the target values are continuous, which is the case in this paper as will be discussed in next sections.

Multi-view Supervised Dictionary Learning

In this section, the formulation for two-view supervised dictionary learning is provided. Extension to more than two views is straightforward. The main assumption is that both views agree on the class labels of all instances in the training set. Let $\mathbf{X}^{(v)} \in \mathbb{R}^{d_1 \times N}$ and $\mathbf{X}^{(w)} \in \mathbb{R}^{d_2 \times N}$ be two views/representations of N training samples with the dimensionalities of d_1 and d_2, respectively. Having these two representations, the main question is how to perform the learning task using the proposed SDL provided in Algorithm 1. There are two approaches, as follows:

Method 1: One approach is to fuse the feature sets from the two views to obtain $\mathbf{X} = \begin{bmatrix}\mathbf{X}^{(v)} \\ \mathbf{X}^{(w)}\end{bmatrix}$, where $\mathbf{X} \in \mathbb{R}^{(d_1+d_2)\times N}$. To learn the supervised dictionary, one needs to use the optimization problem in (5.18). The columns of $\mathbf{U}$, which are the eigenvectors of $\boldsymbol{\Phi} = \mathbf{XHLHX}^{\top}$, construct the dictionary $\mathbf{D} \in \mathbb{R}^{(d_1+d_2)\times l}$, where l is the number of dictionary atoms. Using the formulation given in (5.16), the sparse coefficients $\boldsymbol{\alpha} \in \mathbb{R}^{l\times N}$ can be subsequently computed for both the training and test sets. These coefficients are submitted to a classifier such as SVM for training or classifying an unknown test sample, respectively. As mentioned in the previous subsection, given the data samples $\mathbf{X} \in \mathbb{R}^{(d_1+d_2)\times N}$ and the dictionary $\mathbf{D} \in \mathbb{R}^{(d_1+d_2)\times l}$, the formulation given in (5.16) can be either solved using iterative methods such as the *lasso* or using a closed-form method such as soft-thresholding given in (5.20). The latter has the main advantage that it provides the solution in closed form and hence, in lower computation cost compared to iterative approaches like the *lasso*.

Method 2: The alternative approach is to learn one subdictionary from the data samples in each view. In other words, by replacing $\mathbf{X}^{(v)} \in \mathbb{R}^{d_1 \times N}$ in (5.18) we have

$$\begin{aligned} \max_{\mathbf{U}^{(v)}} \quad & \operatorname{tr}(\mathbf{U}^{(v)^\top} \mathbf{X}^{(v)} \mathbf{HLH} \mathbf{X}^{(v)^\top} \mathbf{U}^{(v)}), \\ \text{s.t.} \quad & \mathbf{U}^{(v)^\top} \mathbf{U}^{(v)} = \mathbf{I}. \end{aligned} \tag{5.21}$$

By computing the corresponding eigenvectors of the largest eigenvalues of $\mathbf{\Phi}^{(v)} = \mathbf{X}^{(v)} \mathbf{HLH} \mathbf{X}^{(v)^\top}$, a subdictionary $\mathbf{D}^{(v)} \in \mathbb{R}^{d_1 \times l_1}$ is obtained, where l_1 is the size of the subdictionary for this view.

Similarly, another subdictionary $\mathbf{D}^{(w)} \in \mathbb{R}^{d_2 \times l_2}$ with the size of l_2 can be computed by replacing $\mathbf{X}^{(w)} \in \mathbb{R}^{d_2 \times N}$ in (5.18), i.e.,

$$\begin{aligned} \max_{\mathbf{U}^{(w)}} \quad & \operatorname{tr}(\mathbf{U}^{(w)^\top} \mathbf{X}^{(w)} \mathbf{HLH} \mathbf{X}^{(w)^\top} \mathbf{U}^{(w)}), \\ \text{s.t.} \quad & \mathbf{U}^{(w)^\top} \mathbf{U}^{(w)} = \mathbf{I} \end{aligned} \tag{5.22}$$

and computing the corresponding eigenvectors of the top eigenvalues of $\mathbf{\Phi}^{(w)} = \mathbf{X}^{(w)} \mathbf{HLH} \mathbf{X}^{(w)^\top}$. By replacing the data samples of each view and their corresponding subdictionaries computed in the previous step in the formulation given in (5.16), the sparse coefficients $\boldsymbol{\alpha}^{(v)} \in \mathbb{R}^{l_1 \times N}$ and $\boldsymbol{\alpha}^{(w)} \in \mathbb{R}^{l_2 \times N}$ can be computed in each view for the training and test samples[4]. Each of these coefficients can be interpreted as the representation of the data samples in the space of the subdictionary of the corresponding view. These coefficients are then fused such that $\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}^{(v)} \\ \boldsymbol{\alpha}^{(w)} \end{bmatrix}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{(l_1+l_2) \times N}$. Fused coefficients $\boldsymbol{\alpha}$ are eventually submitted to a classifier such as SVM for training or classifying an unknown test sample. Algorithms 2 and 3 summarize the computation steps for the two multi-view approaches proposed in this paper.

In the following sections, the relative performance of these two multi-view approaches is shown.

5.3.3 Experimental Study

As follows, the explanation is provided for four approaches in the literature, with which our results are compared. These four approaches are two from dictionary learning and sparse representation literature, one from a recently published paper in multi-view affective speech recognition, and the AVEC 2012 baseline system [188] as described in the following.

[4] The solution can be provided in closed form using (5.20) as mentioned in Method 1.

Algorithm 2 multi-view Supervised Dictionary Learning-Method 1 (MV1)

Input: Training data at multiple views, $\mathbf{X}_{\text{tr}}^{(v)}, v = 1, ..., \text{V}$, test data at multiple views, $\mathbf{X}_{\text{ts}}^{(v)}, v = 1, ..., \text{V}$, kernel matrix of labels $\mathbf{L}$, training data size, N, size of dictionary, l.
Output: Dictionary, $\mathbf{D}$, coefficients for training and test data, $\boldsymbol{\alpha}_{\text{tr}}$ and $\boldsymbol{\alpha}_{\text{ts}}$.

1: $\mathbf{X}_{\text{tr}} = \begin{bmatrix} \mathbf{X}_{\text{tr}}^{(1)} \\ \vdots \\ \mathbf{X}_{\text{tr}}^{(\text{V})} \end{bmatrix}$

2: $\mathbf{X}_{\text{ts}} = \begin{bmatrix} \mathbf{X}_{\text{ts}}^{(1)} \\ \vdots \\ \mathbf{X}_{\text{ts}}^{(\text{V})} \end{bmatrix}$

3: $\mathbf{H} \leftarrow \mathbf{I} - N^{-1}\mathbf{e}\mathbf{e}^\top$

4: $\boldsymbol{\Phi} \leftarrow \mathbf{X}_{\text{tr}}\mathbf{HLH}\mathbf{X}_{\text{tr}}^\top$

5: **Compute Dictionary:** $\mathbf{D} \leftarrow$ eigenvectors of $\boldsymbol{\Phi}$ corresponding to top l eigenvalues

6: **Compute Training Coefficients:** For each data sample $\mathbf{x}_{\text{tr}_i}$ in the fused training set $\mathbf{X}_{\text{tr}}$, use $\alpha_{ij} = S_\lambda([\mathbf{D}^\top \mathbf{x}_{\text{tr}_i}]_j)$, $j = 1, ..., l$ to compute the corresponding coefficient

7: **Compute Test Coefficients:** For each data sample $\mathbf{x}_{\text{ts}_i}$ in the fused test set $\mathbf{X}_{\text{ts}}$, use $\alpha_{ij} = S_\lambda([\mathbf{D}^\top \mathbf{x}_{\text{ts}_i}]_j)$, $j = 1, ..., l$ to compute the corresponding coefficient

Unsupervised k-means

Although *k*-means is known as a clustering approach and hence, an unsupervised technique, in dictionary learning and sparse representation (DLSR) literature, it has been used in both unsupervised and supervised paradigms [65, 207]. In this context, if *k*-means is applied to all training samples on all classes, it is considered as an unsupervised dictionary. However, if the cluster centers are computed on the training samples of each class using *k*-means separately, eventually composed into one dictionary, the dictionary obtained is supervised, and the approach is called supervised *k*-means, which is belonging to one dictionary per class category of SDL approaches mentioned in Section 5.3.1. Supervised *k*-means is designed for discrete labels and it cannot be extended to continuous labels which is the case in affective speech recognition application using dimensional affects. Hence, here, unsupervised *k*-means has been used as one of the dictionary learning approaches to be compared with the proposed approach.

For multi-view learning using *k*-means, the feature sets are first fused and then submitted to the *k*-means for computing the dictionary. The sparse coefficients are learned using (5.16). Since the dictionary is not orthogonal in this case, unlike the proposed approach, (5.16) can be only computed using iterative approaches and it does not have closed-form solution.

Algorithm 3 multi-view Supervised Dictionary Learning-Method 2 (MV2)

Input: Training data at multiple views, $\mathbf{X}_{\text{tr}}^{(v)}, v = 1, ..., \text{V}$, test data at multiple views, $\mathbf{X}_{\text{ts}}^{(v)}, v = 1, ..., \text{V}$, kernel matrix of labels $\mathbf{L}$, training data size, N, size of dictionary, l.

Output: Dictionary, $\mathbf{D}$, coefficients for training and test data, $\boldsymbol{\alpha}_{\text{tr}}$ and $\boldsymbol{\alpha}_{\text{ts}}$.

1: $\mathbf{H} \leftarrow \mathbf{I} - N^{-1}\mathbf{e}\mathbf{e}^\top$

2: **for** $v = 1 \rightarrow \text{V}$ **do**

a: $\boldsymbol{\Phi}^{(v)} \leftarrow \mathbf{X}_{\text{tr}}^{(v)}\mathbf{HLH}\mathbf{X}_{\text{tr}}^{(v)\top}$

b: $\mathbf{D}^{(v)} \leftarrow$ eigenvectors of $\boldsymbol{\Phi}^{(v)}$ corresponding to top l eigenvalues

c: For each data sample $\mathbf{x}_{\text{tr}_i}^{(v)}$ in the training set $\mathbf{X}_{\text{tr}}^{(v)}$, use $\alpha_{ij} = S_\lambda([\mathbf{D}^\top\mathbf{x}_{\text{tr}_i}^{(v)}]_j)$, $j = 1, ..., l$ to compute the corresponding coefficient

d: For each data sample $\mathbf{x}_{\text{ts}_i}^{(v)}$ in the test set $\mathbf{X}_{\text{ts}}^{(v)}$, use $\alpha_{ij} = S_\lambda([\mathbf{D}^\top\mathbf{x}_{\text{ts}_i}^{(v)}]_j)$, $j = 1, ..., l$ to compute the corresponding coefficient

3: **end for**

4: **Compute Dictionary:** $\mathbf{D} \leftarrow \begin{bmatrix} \mathbf{D}^{(1)} & 0 & \cdots & 0 \\ 0 & \mathbf{D}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{D}^{(\text{V})} \end{bmatrix}$

5: **Compute Training Coefficients:** $\boldsymbol{\alpha}_{\text{tr}} \leftarrow \begin{bmatrix} \boldsymbol{\alpha}_{\text{tr}}^{(1)} \\ \vdots \\ \boldsymbol{\alpha}_{\text{tr}}^{(\text{V})} \end{bmatrix}$

6: **Compute Test Coefficients:** $\boldsymbol{\alpha}_{\text{ts}} \leftarrow \begin{bmatrix} \boldsymbol{\alpha}_{\text{ts}}^{(1)} \\ \vdots \\ \boldsymbol{\alpha}_{\text{ts}}^{(\text{V})} \end{bmatrix}$

Discriminative K-SVD

To provide a comparison with the supervised dictionary learning (SDL) approaches in the literature, as mentioned in Section 5.3.1, not all the proposed SDL methods in the literature are extendible to continuous labels. For example, all of the SDL methods in category 1 mentioned in Section 5.3.1, i.e., one dictionary per class category, need discrete class labels and none of them can be applied to continuous labels. Among the SDL approaches in the literature, we have chosen the discriminative K-SVD (DK-SVD) [237] approach that jointly learns the dictionary and a linear classifier in one optimization problem. Although DK-SVD was originally proposed for classification problem, i.e., for discrete labels, it can be easily extended to regression problems (for continuous labels). It is sufficient to replace the discrete labels in the formulation provided in [237] with continuous labels, all other steps remain unchanged.

To implement multi-view DK-SVD, the same as multi-view *k*-means, the features from single views are fused and then submitted to the DK-SVD formulation provided in [237].

Cross-Modal Factor Analysis

The proposed multi-view SDL approach in this paper is a supervised multi-view technique as the class labels are included in the learning process. There are, however, unsupervised approaches in the literature that perform multi-view analysis by including the correlation among the views into the learning process without taking into account the class labels. Cross-modal factor analysis (CFA) [112] is one of these approaches, which has recently been introduced in the context of multi-view affective speech recognition [214]. CFA is an unsupervised approach that includes the relationship between the two views by minimizing the ℓ_2 norm distance between the projected points into two orthogonal subspaces.

Subsequently, the projected data points into the coupled subspaces are computed and concatenated to jointly represent the data. They are eventually submitted to a regressor for its training using the training set, and subsequently predicting the dimension of an unknown emotion. Unlike other approaches discussed in this paper, CFA does not lead to a sparse representation.

AVEC 2012 Baseline System

The AVEC 2012 baseline system [188] is comprising of baseline features submitted to support vector machines regression (SVR). Here, the original baseline feature set is used with a dimensionality of 1841 features, whereas in previous three approaches, the dimensionality is determined by the dictionary size (in unsupervised *k*-means and DK-SVD) or the number of components in the jointly learned subspaces (in CFA), which is far less than the original feature set size in our experiments (maximum 64).

Setup

Two feature sets described above have been used, i.e., SED and baseline features, as the two views v and w for an affective speech recognition system based on the multi-view SDL proposed earlier in this paper. Hence, the two views are $\mathbf{X}^{(v)} \in \mathbb{R}^{400\times N}$ and $\mathbf{X}^{(w)} \in \mathbb{R}^{1841\times N}$, where N is 10806 in the training set (which is used for both training and tuning the parameters) and 9312 in the development set (which serves as the test set) for the FCSC part of the dataset in the experiments.

There are four dimensional affects, i.e., arousal, expectation, power, and Valance, as the continuous response variables to be predicted. Hence, a regression model is to be deployed in the system. The *lasso* regressor and its GLMNET [5] implementation are used in all approaches except for DK-SVD that learns its own linear coefficients and AVEC 2012 baseline system that uses SVR. The sparsity parameter of the lasso has been optimized over the training set by a 10-fold cross validation. As for the SVR, a linear-kernel is used in the experiments and the trade-off parameter (C^*) of the SVR is tuned by a line search over the set of values of $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, and by 5-fold cross validation on the training set.

An RBF kernel is used over the response variable in each dimension, which serves as the kernel over the target values ($\mathbf{L}$) to compute $\mathbf{\Phi}$ in Algorithms 2 and 3. The kernel width of the RBF kernel has been set by using a self-tuning approach similar to what is explained in [236], i.e., $\sigma_i = \frac{1}{N_{\text{train}}} \sum_{j \neq i} d(y_i, y_j)$, which is the average (Euclidean) distance between a response variable and all others. The training set is used to compute the dictionary. The optimal value of the regularization parameter in soft thresholding (λ^*) for the proposed multi-view dictionary learning methods, which controls the level of sparsity, has been computed by 10-fold cross-validation on the training set. The λ^* is then used to compute the coefficients for both training and test sets[6].

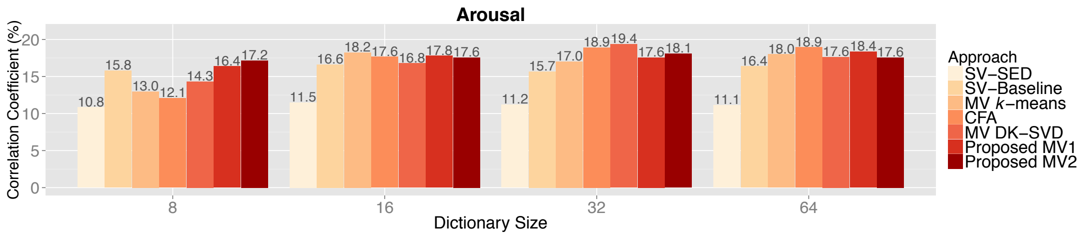
In all experiments, the data in each view is normalized such that each feature is mapped to the range of [0,1]. As suggested in [188], performance is evaluated using Pearson's correlation coefficient (CC) for each session. The correlation between the predicted and actual values is calculated for each session. However, since sessions are of different lengths, the contribution of each session in the total correlation should be different. Therefore, to calculate the overall correlation coefficient, we have used the weighted average of session correlations, where sessions' lengths are used as for the weights.

Results and Discussions

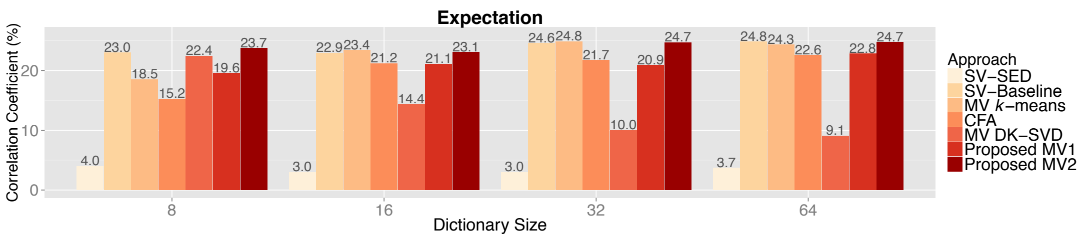
The correlation coefficients (CC) for HSIC-based SDL at single view (Algorithm 1) and also for the proposed multi-view approach to affective speech recognition (Algorithms 2 and 3) and rival multi-view approaches computed over the two feature sets, i.e., SED and baseline features, are reported in Figure 5.14 for the arousal, expectation, power, and valence dimensions at four dictionary sizes, i.e., 8, 16, 32, and 64. The average over all four dimensions of learning time (including the time required to learn the dictionary and coefficients, the tuning time for the sparsity coefficient of the regressor, and also the time for training the regressor) as well as recall

[5] http://www-stat.stanford.edu/∼tibs/glmnet-matlab/.

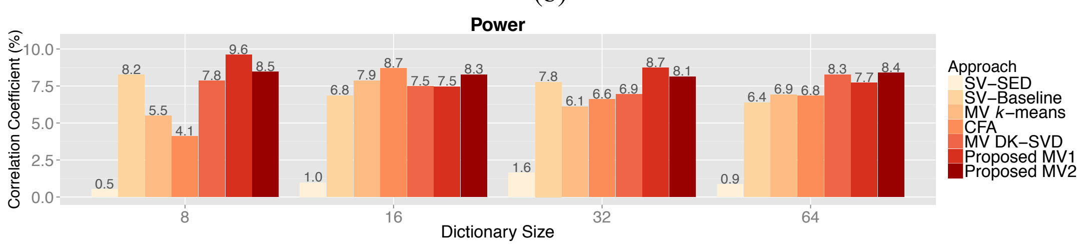
[6] One λ^* is computed for each data point in the training set. However, the averaged λ^* over the whole training set is used to compute the coefficients on the training and test sets as it yields better generalization.



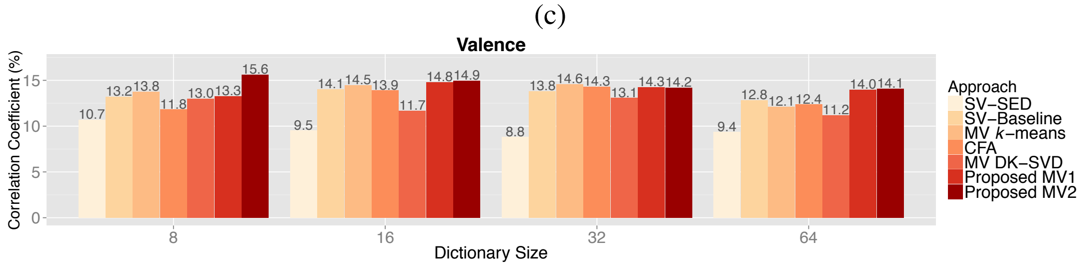
(a)



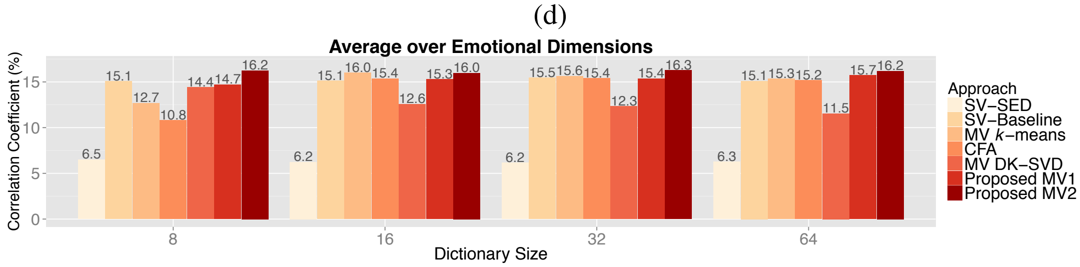
(b)



(c)



(d)



(e)

Figure 5.14: The percentage of correlation coefficient (CC) based on single-view (SV) and multi-view (MV) learning approaches. MV1 and MV2 are the multi-view SDL techniques based on Algorithms 2 and 3, respectively as discussed in Section 5.3.2. The results are shown at four different dictionary sizes for (a) arousal, (b) expectation, (c) power, (d) valence, and (e) average over all dimensional affects.

Table 5.6: The average learning time (including the time required for learning the dictionary and the coefficients, tuning the sparsity parameter for the lasso, and eventually training the using tuned parameters) and recall time (in seconds) for the single-view and multi-view approaches. The computation time is averaged over all the dimensional affects for each method. The results are reported for four dictionary sizes 8, 16, 32, and 64.

		Dictionary Size			
	Approach	**8**	**16**	**32**	**64**
Learning Time	SV-SED	35	42	73	199
	SV-Baseline	82	98	206	490
	MV *k*-means	149	267	731	2168
	CFA	139	139	139	139
	MV DK-SVD	359	736	1400	4965
	Proposed MV1	104	125	192	384
	Proposed MV2	78	88	119	302
Recall Time	SV-SED	0.022	0.026	0.027	0.038
	SV-Baseline	0.037	0.042	0.043	0.064
	MV *k*-means	4.687	14.786	51.829	170.346
	CFA	0.034	0.022	0.028	0.037
	MV DK-SVD	0.673	1.056	2.334	7.391
	Proposed MV1	0.047	0.048	0.064	0.069
	Proposed MV2	0.061	0.060	0.061	0.065

Table 5.7: The percentage of correlation coefficient (CC), learning, and recall time (in seconds) for the AVEC 2012 baseline system using the baseline features and a support vector machine regression (SVR) with linear kernel.

	Arousal	**Expectation**	**Power**	**Valence**	**Average**
Correlation Coefficient	19.9	23.1	8.7	7.5	14.8
Learning Time	15684	27941	14676	20743	19761
Recall Time	497	959	499	706	665

(test) time are provided in Table 5.6. Since there is no dictionary associated with the AVEC 2012 baseline system, the results related to this approach are separately provided in Table 5.7. The p values for the statistical test of significance (paired t-test) performed pairwise between the proposed multi-view approaches and all single view or rival approaches are reported in Table 5.8.

As can be seen in Figure. 5.14, both proposed multi-view approaches (MV1 and MV2) benefit from the complementary information in two-view features sets. The performance of the single-view system based on the SED is usually inferior to the one based on the baseline feature set. However, combining these two representations using one of the proposed multi-view approaches discussed earlier leads to higher correlation coefficients in all dimensions (except for MV1 in expectation dimension). The results of statistical significance test (Table 5.8) show that both MV1 and MV2 significantly outperform ($p < 0.05$) single view method based on SED features. Moreover, MV2 significantly outperforms the other single view method, which is using baseline features.

For the purpose of comparing the proposed multi-view SDL methods with the AVEC 2012 baseline system, if we take the average of correlation coefficient over all dimensions and dictionary sizes, MV1 and MV2 achieve an average performance of 15.27% and 16.17%, respectively, whereas the average correlation coefficient over all four dimensions for the AVEC 2012 baseline system is 14.8%, which is less than (although not significant according to Table 5.8) the performance of the proposed methods. Also, since original baseline features, i.e., 1841 features, are used in the AVEC 2012 baseline system, the dimensionality is much higher than the dictionary learning approaches (maximum 64). Consequently, the computational time for both learning and recalling are much longer than all other approaches. For example, the average recall time over all dimensions for the AVEC 2012 baseline system (665 s) is more than 10000 times longer than the same for the proposed MV1 (0.057) and MV2 (0.062 s).

Furthermore, the proposed MV2 significantly (see Table 5.8) outperforms other multi-view approaches in the literature. Also, the performance of the proposed MV1 is significantly better than MV DK-SVD. Supervised multi-view methods, i.e., multi-view DK-SVD, MV1, and MV2 particularly benefit from the information in target values' information (dimensional affects) at small dictionary size as can be observed from the results at the dictionary size of 8 in Figure. 5.14. For example, for power dimensional affect, MV1 performs about twice as good as the unsupervised multi-view techniques, i.e., k-means and CFA. By increasing the dictionary size, however, the unsupervised multi-view approaches can capture the underlying correlation among the single view feature sets, hence their performance approaches those of the supervised multi-view techniques. Nevertheless, the main advantage of better performance at small dictionary sizes is much lower computational cost, as increasing the number of dictionary atoms also increases the computational time. On the other hand, between the two supervised approaches, while the proposed multi-view approaches provide a closed-form solution for both the dictio-

Table 5.8: Tests of statistical significance (paired t-test) between proposed multi-view methods (MV1 or MV2) and single view or rival multi-view approaches. p-values are shown for the proposed MV methods vs. the single view or rival approach. (* denotes $p < 0.05$; ** denotes $p < 0.01$; *** denotes $p < 0.001$.)

	SV-SED	**SV-Baseline**	**MV *k*-means**	**CFA**	**MV DK-SVD**	**Baseline**
MV1	0.000^{***}	0.853	0.495	0.054	0.035^{*}	0.641
MV2	0.000^{***}	0.000^{***}	0.010^{*}	0.007^{**}	0.016^{*}	0.164

nary and coefficients, multi-view DK-SVD optimization problem is nonconvex and the solution has to be performed iteratively and alternately for the dictionary and coefficients [237] using an iterative algorithm such as K-SVD [1]. This has two main disadvantages, first, it increases the computation time, and second, the algorithm may get stock in a local minimum solution. The latter disadvantage of DK-SVD algorithm explains its poor performance in expectation dimension for the dictionary sizes of 16, 32, and 64. Moreover, in average, the performance of DK-SVD is far behind the proposed MV1 and MV2. Not to mention that it is learning time is the longest after AVEC 2012 baseline system, as tuning its parameters is very time consuming, and makes this approach unsuitable in the applications where online learning is required.

In terms of the complexity of methods, the proposed multi-view approaches are the least complex as their solution is closed form for both the dictionary and coefficients. Although learning the dictionary and coefficients does not have to be done iteratively and alternately for the MV k-means method, neither the dictionary nor the coefficients can be learned in closed form, which makes both learning and recalling time for this method relatively long (see Table 5.6). As can be seen in Table 5.6, the proposed MV1 and MV2 are computationally much more efficient than the other two dictionary-based multi-view approaches, i.e., k-means and DK-SVD. Although CFA also offers a closed-form solution using singular value decomposition, unlike MV1 and MV2, it does not lead to a sparse representation in the subspaces.

Both CFA and proposed multi-view approaches can be kernelized. The formulation for the kernelized CFA has been provided in [214]. A kernelized version of HSIC-based SDL was proposed in [65]. The extension to multi-view learning is straightforward and leads to similar algorithms as in Algorithms 2 and 3. However, the kernelized version of the proposed multi-view approach will lead to a sparse representation, which is an advantage for the approach compared to the kernelized CFA. The proposed MV1 and MV2 approaches can be easily extended to more than two views as shown in Algorithms 2 and 3. This is not the case for the extension of the CFA to more than two views as the correlation between every two views has to be computed pairwise using an optimization problem given in [214]. However, this may not lead to unique solutions for the subspaces.

Considering that MV1 and MV2 achieve an average correlation coefficient over all dictionary sizes and dimensions of 15.27% and 16.17%, respectively reveals higher performance of MV2 compared to MV1 in average. If we also take into account the computation time, that is learning time for MV2 is faster than MV1, MV2 seems to be the more favorable of the two.

As a final remark, it is worth to mention that MV2 learns the dictionary and coefficients in the two views independently, and only fuses the features in the space of leaned dictionaries at the final stage. This is expected to be useful when the two views are independent or not very much correlated. If this is not the case, learning the dictionary in a fused space of two views might be beneficial, as the dictionary learned can share the common properties of both views. This can be especially useful for small dictionary sizes.

5.4 Conclusion

Variable selection for affective speech recognition is firstly discussed in this chapter. For that purpose, a wide range of dimensionality reduction algorithms are used. The experiments are run in the framework of VAM dataset, as well Interspeech 2012 and 2013 paralinguistic challenges. Result of the performed experiments are presented in the form of a short list of speech features, compared to the list of commonly used features, that could result in competitive prediction accuracy. In doing so, low-level descriptors are distinguished from functionals, and their relative importance for the recognition problem is studied independently.

The second topic discussed in this chapter is max-dependence regression (MDR) as a learning algorithm. The objective of MDR is to estimate parameters of the linear model with the intention of maximizing the dependence between predicted and actual values of the response variable. MDR is compared to the state-of-the-art support vector regression (SVR) under different scenarios and in different experiments, and it is shown that MDR outperform SVR from accuracy point of view, and that the former is far less computationally expensive that the latter.

Finally, discussed in this chapter is the application of dictionary learning to affective speech recognition. It is shown that given an affective speech dataset, using a small number of bases that could capture the essences of the dataset could improve generalization capabilities of a statistical model. Furthermore, using the dictionary learning approach of choice, multi-view learning for dictionaries is proposed, which is meant to combine different datasets of different nature. It is shown that the proposed multi-view approaches outperform the state-of-the-art single view and multi-view approaches.

Chapter 6

Conclusion and Future Works

We summarize this thesis in this chapter by referring to the four questions that we asked earlier, and by summarizing our proposed answer to each question.

6.1 Speech Features

First question that we asked in the first chapter was *What set of speech features can carry its affective properties?* We first addressed this question in Section 4.1, where we proposed spectral energy distribution (SED) as a set of speech low-level descriptors. We showed that SED is composed of a set of components each of which is defined over an interval of the frequency domain, and that components are normalized in a way that their variation would matter at later stages for statistical learning, rather than their magnitude. Later, we put SED into practice by using them for classifying binarized dimensional affect including arousal, expectancy, power, and valence, and we showed that very few SED based features surpass informativeness of thousands of features commonly used in the literature. In other words, we showed that 15 SED based features resulted in higher average prediction accuracy than other studies on the same dataset. This is particularly valuable, since those works have used longer feature vectors, i.e., ranging from hundreds to thousands of features.

Then, we applied SED to modeling of continuous dimensional affect in the framework of the second audio/visual emotion challenge, and we noticed that the proposed set of features complement the baseline features of the challenge, however they may not be sufficient by their own.

We also examined the proposed set of features for modeling paralinguistic qualities of speech such as personality, likability, and intelligibility, and we discovered that SED features offer a more concise and more accurate model for certain personality dimensions, as well as for likability, and for the most of other dimensions they give higher results when they are used alongside other features; in some cases, the combination of those features did not result in higher accuracy.

Furthermore, we proposed a measure for quantifying speed of temporal changes in speech, as a statistical functional defined based on SED in Section 4.2. The proposed functional, named dynamicity, takes SED as a probability distribution function, as is used to measure the amount of changes from one frame of the speech signal to the next. We used dynamicity features for classifying binarized paralinguistic qualities of speech such as affect, conflict, and autism, and we showed that the very short vector of dynamicity features offers high separability for the aforementioned paralinguistic qualities.

Additionally, we performed set of dimensionality reduction experiments in Section 5.1. First, we discussed the problem of dimensionality reduction, and we explained why we favor selection algorithms over projection methods, since the latter offers a mixture of original features, due to which we lose interpretability. Furthermore, by choosing the mix together audio features, we would still need to extract all the existing features in the feature vector, which in return may not be in the favor of lower computational complexity at the extraction stage. In contrast, selection algorithms allow us to lower the computational complexity of feature extraction, and the resulting set of features would be interpretable in the sense that we would know what speech features would be useful for extracting affective contents of speech, and which would not.

In the first set of experiments, where the objective was to model continuous affective dimensions using the spontaneous affective dataset VAM, we compared four different dimensionality reduction algorithms to find the best set of features for predicting arousal, dominance, and valence. The selected set of algorithms covered all types of dimensionality reduction algorithms: wrappers, filters, and projection methods, as well as both supervised and unsupervised reduction algorithms. Through that experiment we showed that combining different sets of features, even in spite of utilizing ranges of dimensionality reduction algorithms, does not necessarily result in higher prediction accuracy than the cases in which we use only one of the feature sets, and that by fixing the dimensionality of the destination feature space, one of the two feature sets would result in higher accuracy.

In the second set of experiments, we considered a wide range of paralinguistic qualities for which we performed dimensionality reduction. Those qualities are arousal and valence from affective dimensions, autism, conflict, intelligibility, likability, and the five personality dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism. We based those experiments on the paralinguistic challenges as conducted in the Interspeech 2012 and 2013. The baseline feature vector provided for those challenges comprises more than 6000 features. By performing variable selection for the aforementioned set of paralinguistic qualities, we showed that how by using less than 5% of those features in most of the cases, we could achieve higher prediction accuracy.

Furthermore, the performed set of experiment is of a particular value, as firstly they list the selected set of features, and secondly, they distinguish selected low-level descriptors and statistical functionals. The advantage of this analysis is twofold: 1) it provides a great amount of insight as to what characteristics of speech are relevant to each of the paralinguistic qualities, by taking into account every detail of the extracted set of features, and 2) they can save us a considerable amount of computational expenses at feature extraction, and at parameter estimation stages. At the feature extraction stage, firstly because we may not need to extract all the long list of low-level descriptors, and secondly, because we may not need to compute all the long list of functionals for contours of the extracted set of features. On the other hand, at the parameter estimation for statistical learning, the reduced set of speech features would allow us to considerably cut computational expenses of learning algorithms, whose complexity is in most cases at least quadratic in the length of the feature vector.

6.2 Individual Differences

The second question that we asked in the first chapter was *Does an answer to the first question depend on individual human beings?* To answer to this question, we proposed the concept of spectral emotion profile (SEP) in Section 4.3. SEP is to formalize individual differences, as well as any other context (e.g. cultural, gender-related) that could make a difference in the expression of affect, by focusing on the variations that those differences could cause in the spectrum of the speech signal. In other words, since different spectral intervals have uneven contributions to capturing affective contents of speech, and since these two factors depend on the characteristics of a speaker, SEP is proposed to take account of all those differences.

Therefore, we put the concept of SEP into practice in two different contexts: gender-specific and individual profiles. For those experiments, we considered both categorical and dimensional representations of affect. For the categorical affect, we considered anger, boredom, disgust, fear, happiness, neutral, and sadness. And, for the dimensional affect, we considered activation, dominance, and valence. The result of applying SEP for different genders proved that the effective choice of spectral intervals is different in the case of the female speakers, from the case of male speakers, and both of those two cases are different than the case of gender-independent spectral profile. Furthermore, for the case of individuals' spectral emotion profiles, we noticed how drastically different those profiles could be from one person to another. We also showed how the number of required SED components can considerably reduce when then identity of a speaker is provided.

6.3 Statistical Learning

The next question that we asked in the first chapter was *What statistical learning algorithms are suitable for affective speech recognition?* In the literature of affective computing, correlation coefficient is prevalently used as a measure of goodness of fit for regression models for modeling continuous dimensional affect. That is to say, between two models, the one that results in a higher correlation coefficient is favored over the other one. However, the literature of this problem before this research did not show any adaptation to this criterion. Proposed in this thesis (Section 5.2) is a learning algorithm that is developed having this very objective in mind, i.e., to maximize the dependence between predicted and actual values of the response variable. The proposed algorithm is called max-dependence regression (MDR) and is developed to maximize Hilbert-Schmidt independence criterion, which is a generic sense of independence. Therefore, a great advantage of MDR is that it can capture nonlinear dependencies.

We have put MDR into practice in several different experiments. In one experiment we used MDR to estimate coefficient of the linear model for modeling arousal, expectancy, power, and valence dimensions in two different time granularities: fully-continuous and word-level. In this experiment, we used support vector regression (SVR) as a state-of-the-art learning algorithm for the linear model as the base of our comparisons. We used prediction error and correlation coefficient as the measures of our comparisons. Results of this experiment showed that MDR outperforms SVR for all eight cases, i.e., four affective dimensions and two time granularities. According to this experiment, MDR resulted in more than 50% improvement over SVR in six out of eight cases. On the other hand, by taking the processing time as a measure of computational complexity, MDR turned out to be much less computationally expensive than SVR. MDR showed to be in average 17 and 22 times faster than SVR in terms of the training time for the fully-continuous and word-level recognition tasks, respectively. And, the recall time of MDR was 60 time faster than that of the SVR.

In another experiment, we used MDR for modeling activation, dominance, and valence for spontaneous affective speech. Again, we used prediction error and correlation coefficient to compare MDR with SVR for the same learning task. We performed two different hypothesis testing, one by k-fold cross validation, and another by leave-one-subject-out cross validation. For both types of testing, for activation and dominance dimensions MDR and SVR resulted in similar accuracies, however, for the valence dimension, SVR outperformed MDR.

To further investigate capabilities of MDR, we used it in two toy problems. For each of those problems, we considered three factors of variation: size of data, nonlinearity, and randomness. By varying those factors, we generated 300 different regression problems, 150 for each of the toy problems. Accordingly, we used MDR and SVR to train models for each of the individual problems. Comparing prediction accuracies of MDR and SVR for the 300 problems showed that MDR performs better in 256 cases, which translates into more than 85% of the cases. On the other hand, although the two algorithms showed similar recall time for the studies cases, the average training time of MDR showed to be 25 and 13 times faster than that of the SVR.

6.4 Sparse Representation

The last question that we asked earlier in this work was *In the presence of multitude of learning patterns, what combination of those can concisely capture their essence?* To answer to this question, we used the concept of dictionary learning and sparse representation. More precisely, we used a supervised dictionary learning (SDL) algorithm based on the Hilbert-Schmidt independence criterion to address this problem. As a result of SDL, we were able to concisely describe affective speech datasets using very few bases. Before the current research, DLSR has not been used in the literature of affective speech recognition. Furthermore, we used a multi-view supervised dictionary learning to fuse feature vectors of different nature. For this purpose, two different types of fusion are seen: 1) fusion at bases level, and 2) fusion as feature level.

Based on the aforementioned multi-view supervised dictionary learning algorithms, we run experiments in the framework of audio/visual emotion challenge 2012. The number of training instances in the data set is close to 20000. Therefore, the objective of those experiments was to find a basis with a lower dimensionality than that of the original dataset, and to examine the learned bases by learning a regression model on each, and to compare the prediction accuracy of each resulting regression model. For this purpose, different dictionary sizes are used. We showed that using only 8 bases, we could represent the dataset, and yet obtain higher prediction accuracy that the case where all the data points where used for estimating the coefficient of the regression model.

Furthermore, we compared the HSCI-based dictionary learning approaches with the state-of-the-art dictionary learning approaches such as K-means dictionary learning, and we showed that HSIC-based approaches result in higher prediction accuracy for equal dimensionality of bases, and that they are far less computationally expensive to learn.

6.5 Future Works

The application of spectral emotion profile for personalizing affective speech recognition is considered in this work. As extra contexts, gender of subjects, as well as their identity is used. Cultural background of the speakers could be also used for this purpose. Realization of this objective, however, would depend on the availability of datasets that could include speech samples of individuals from different backgrounds, which is not accommodated by the already existing datasets. Therefore, as a future direction of this research, we would like to propose the collection of such affective speech dataset, and the extension of the application of spectral emotion profile to cultural backgrounds.

As discussed in this thesis, variable selection plays a crucial role in the success of a statistical model for affective speech. On the other hand, the application of max-dependence regression (MDR) as a learning algorithm for regression model is proposed in this thesis. As a future direction of this research, we would like to propose the design of sparse max-dependence regression (SMDR). That is, an algorithm that would oblige sparsity of the regression coefficients by introducing a penalty function, possibly ℓ_1 norm, to the optimization criterion. Therefore, the resulting algorithm would perform variable selection together with estimating the regression coefficients. The advantage of such an algorithms would be that variable selection and parameter estimation are done at the same time and towards reaching a single goal.

As another extension of max-dependence regression, we would like to propose the application of the proposed criterion for this algorithm to more elaborate statistical models, such as deep neural networks. That is, to estimate parameters of a deep neural network by maximizing the Hilbert-Schmidt independence criterion as a generic measure of independence, instead of minimizing the prediction error.

A useful application of affective speech recognition is for improving the recognition accuracy of automatic speech recognition. That is, to conduct the recognition task of phonemes based on the affective contents of speech, in order to account for possible distortions that might have been caused by extreme affective expressions. Realization of this goal would depend on the availability of a automatic speech recognition dataset that could account for the affective contents of phonemes, which is not accommodated by the already existing datasets. Therefore, as a future direction of this research, we would like to propose the collection of such a dataset, and the application of affective speech recognition to improve on the accuracy of the already existing automatic speech recognition algorithms.

Another interesting application of affective speech recognition is the inverse problem, that is to synthesize speech that is not affectively neutral. Based on the current research, the analysis of relevant variables to the expression of affect could be efficiently used towards the realization of this goal. This goal might be reached firstly by analyzing what features, and in what manner, have to be changed in order to depart from the neutral affective state, and then apply those changes to neutrally expressed phonemes, and smooth the transition from one phoneme to another, in order to account for the modification that would have been done in order to emotionally bias a synthesized speech sample.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, Nov. 2006.

[2] Senaka Amarakeerthi, Tin Lay Nwe, Liyanage C De Silva, and Michael Cohen. Emotion classification using inter-and intra-subband energy variation. In *INTERSPEECH*, pages 1569–1572, 2011.

[3] Massih R. Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views - an application to multilingual text categorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 28–36, 2009.

[4] Gouzhen An, David-Guy Brizan, and Andrew Rosenberg. Detecting laughter and filled pauses using syllable-based features. In *INTERSPEECH*, pages 178–181, 2013.

[5] Gopala Krishna Anumanchipalli, Hugo Meinedo, Miguel Bugalho, Isabel Trancoso, Luís C Oliveira, and Alan W Black. Text-dependent pathological voice detection. In *INTERSPEECH*, 2012.

[6] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[7] Meysam Asgari, Alireza Bayestehtashk, and Izhak Shafran. Robust and accurate features for detecting and diagnosing autism spectrum disorders. In *INTERSPEECH*, pages 191–194, 2013.

[8] Yazid Attabi and Pierre Dumouchel. Anchor models and wccn normalization for speaker trait classification. In *INTERSPEECH*, 2012.

[9] Kartik Audhkhasi, Angeliki Metallinou, Ming Li, and Shrikanth Narayanan. Speaker personality classification using systems based on acoustic-lexical cues and an optimal tree-structured bayesian network. In *INTERSPEECH*, 2012.

[10] Eric Bair, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101:119–137, 2006.

[11] Tadas Baltrušaitis, Ntombikayise Banda, and Peter Robinson. Dimensional affect recognition using continuous conditional random fields. In *IEEE Conference on Automatic Face and Gesture Recognition*, 2013.

[12] Roberto Barra Chicote, Fernando Fernández Martínez, Lebai Lutfi, Syaheerah Binti, Juan Manuel Lucas Cuesta, Javier Macías Guarasa, Juan Manuel Montero Martínez, Rubén San Segundo Hernández, and José Manuel Pardo Muñoz. Acoustic emotion recognition using dynamic bayesian networks and multi-space distributions. In *INTERSPEECH*. ISCA, 2009.

[13] Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357 – 1371, 2011.

[14] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 8 edition, 2009.

[15] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[16] Tobias Bocklet, Georg Stemmer, Viktor Zeissler, and Elmar Nöth. Age and gender recognition based on multiple systems-early vs. late fusion. In *INTERSPEECH*, pages 2830–2833, 2010.

[17] Daniel Bone, Theodora Chaspari, Kartik Audhkhasi, James Gibson, Andreas Tsiartas, Maarten Van Segbroeck, Ming Li, Sungbok Lee, and Shrikanth Narayanan. Classifying language-related developmental disorders from speech cues: the promise and the potential confounds. In *INTERSPEECH*, pages 182–186, 2013.

[18] Elif Bozkurt, Engin Erzin, Cigdem Eroglu Erdem, and A Tanju Erdem. Improving automatic emotion recognition from speech signals. In *INTERSPEECH*, pages 324–327, 2009.

[19] Raymond Brueckner and Björn Schuller. Likability classification-a not so deep neural network approach. In *INTERSPEECH*, 2012.

[20] Harm Buisman and Eric O Postma. The log-gabor method: speech classification using spectrogram image analysis. In *INTERSPEECH*, 2012.

[21] F Burkhardt, A Paeschke, M Rolfes, W F Sendlmeier, and B Weiss. A database of german emotional speech. In *in Proceedings of Interspeech, Lissabon*, pages 3–6, 2005.

[22] Felix Burkhardt, Björn Schuller, Benjamin Weiss, and Felix Weninger. " would you buy a car from me?"-on the likability of telephone voices. In *INTERSPEECH*, pages 1557–1560, 2011.

[23] C. Busso, Sungbok Lee, and S. Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(4):582 –596, may 2009.

[24] Claudia Caffi and Richard W. Janney. Toward a pragmatics of emotive communication. *Journal of Pragmatics*, 22(3 - 4):325 – 373, 1994.

[25] Ricardo Calix, Mehdi Khazaeli, Leili Javadpour, and Gerald Knapp. Dimensionality reduction and classification analysis on the audio section of the semaine database. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 323–331. Springer Berlin / Heidelberg, 2011.

[26] George Caridakis, Lori Malatesta, Loic Kessous, Noam Amir, Amaryllis Raouzaiou, and Kostas Karpouzis. Modeling naturalistic affective states via facial and vocal expressions recognition. In *Proceedings of the 8th international conference on Multimodal interfaces*, ICMI '06, pages 146–154, New York, NY, USA, 2006. ACM.

[27] S. Casale, A. Russo, and S. Serrano. Analysis of robustness of attributes selection applied to speech emotion recognition. In *18th European Signal Processing Conference (EUSIPCO-2010)*, pages 1174–1178, 2010.

[28] Ling Cen, Zhu Yu, and Ming Dong. Speech emotion recognition system based on l1 regularized linear regression and decision fusion. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 332–340. Springer Berlin / Heidelberg, 2011.

[29] Suryannarayana Chandaka, Amitava Chatterjee, and Sugata Munshi. Support vector machines employing cross-correlation for emotional speech recognition. *Measurement*, 42(4):611 – 618, 2009.

[30] Guillaume Chanel, Joep J. M. Kierkels, Mohammad Soleymani, and Thierry Pun. Short-term emotion assessment in a recall paradigm. *Int. J. Hum.-Comput. Stud.*, 67(8):607–627, August 2009.

[31] Clément Chastagnol and Laurence Devillers. Personality traits detection using a parallelized modified sffs algorithm. In *INTERSPEECH*, volume 15, page 16, 2012.

[32] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487 – 503, 2008.

[33] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32 –80, jan 2001.

[34] Albert Cruz, Bir Bhanu, and Songfan Yang. A psychologically-inspired match-score fusion model for video-based facial expression recognition. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 341–350. Springer Berlin / Heidelberg, 2011.

[35] Albert C. Cruz, Bir Bhanu, and Ninad Thakoor. Facial emotion recognition with expression energy. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 457–464, New York, NY, USA, 2012. ACM.

[36] Nicholas Cummins, Julien Epps, and Jia Min Karen Kua. A comparison of classification paradigms for speaker likeability determination. In *INTERSPEECH*, 2012.

[37] Mohamed Dahmane and Jean Meunier. Continuous emotion recognition using gabor energy filters. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 351–358. Springer Berlin / Heidelberg, 2011.

[38] Charles Darwin, Paul Ekman, and Phillip Prodger. *The Expression of the Emotions in Man and Animals*. Oxford University Press, 2002.

[39] David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

[40] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. K. J. Heylen. The sensitive artificial listner: an induction technique for generating emotionally coloured conversation. In *LREC Workshop on Corpora for Research on Emotion and Affect, Marrakech, Marokko*, pages 1–4, Paris, France, 2008. ELRA.

[41] Pierre Dumouchel, Najim Dehak, Yazid Attabi, Reda Dehak, and Narjes Boufaden. Cepstral and long-term features for emotion recognition. In *Interspeech*, pages 344–347, 2009.

[42] Paul Ekman. *Basic Emotions*. Sussex, U.K.: John Wiley & Sons, Ltd, 1999.

[43] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, Dec. 2006.

[44] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1600–1607, 2012.

[45] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2797, 2009.

[46] H.P. Espinosa, C.A.R. Garcia, and L.V. Pineda. Features selection for primitives estimation on emotional speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5138 –5141, march 2010.

[47] H.P. Espinosa, C.A.R. Garcia, and L.V. Pineda. Bilingual acoustic feature selection for emotion estimation using a 3d continuous model. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 786 –791, march 2011.

[48] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie. Online emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3:7–19, 2010.

[49] Florian Eyben, Anton Batliner, Björn Schuller, Dino Seppi, and Stefan Steidl. Cross-corpus classification of realistic emotions – some pilot experiments. In *Proc. of Third International Workshop on EMOTION (satellite of LREC): CORPORA FOR RESEARCH ON EMOTION AND AFFECT*, pages 77–82, May 2010.

[50] A.K. Farahat, A. Ghodsi, and M.S. Kamel. An efficient greedy method for unsupervised feature selection. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 161 –170, dec. 2011.

[51] Michael Feld, Felix Burkhardt, and Christian A Müller. Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services. In *INTERSPEECH*, pages 2834–2837, 2010.

[52] E. Fersini, E. Messina, and F. Archetti. Emotional states in judicial courtrooms: An experimental investigation. *Speech Communication*, 54(1):11 – 22, 2012.

[53] P. Fewzee and F. Karray. Dimensionality reduction for emotional speech recognition. In *Proceedings of IEEE SocialCom 2012*, pages 532–537, September 2012.

[54] P. Fewzee and F. Karray. Elastic net for paralinguistic speech recognition. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI 12, pages 509–516, New York, NY, USA, 2012. ACM.

[55] P. Fewzee and F. Karray. Emotional speech: A spectral analysis. In *Proceedings of Interspeech 2012*, pages 2238–2241, September 2012.

[56] P. Fewzee and F. Karray. Continuous emotion recognition: Another look at the regression problem. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 197 – 202, September 2013.

[57] Johnny Fontaine, KR SCHERER, EB ROESCH, and PC ELLSWORTH. The world of emotions is not two-dimensional. *PSYCHOLOGICAL SCIENCE*, 18(12):1050–1057, 2007.

[58] N. Fragopanagos and J.G. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389 – 405, 2005.

[59] J.H. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[60] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto. Localizing objects with smart dictionaries. In *Proceedings of the* 10^{th} *European Conference on Computer Vision (ECCV): Part I*, pages 179–192, 2008.

[61] Rok Gajšek, Janez Žibert, Tadej Justin, Vitomir Štruc, Boštjan Vesnicer, and France Mihelič. Gender and affect recognition based on gmm and gmm-ubm modeling with relevance map estimation. In *INTERSPEECH*, 2010.

[62] M. J. Gangeh, L. Sørensen, S. B. Shaker, M. S. Kamel, M. de Bruijne, and M. Loog. A texton-based approach for the classification of lung parenchyma in CT images. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 595–602, Berlin, Heidelberg, 2010. Springer-Verlag.

[63] Mehrdad J Gangeh, Pouria Fewzee, Ali Ghodsi, Mohamed S Kamel, and Fakhri Karray. Multiview supervised dictionary learning in speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(6):1056–1068, June 2014.

[64] Mehrdad J. Gangeh, Ali Ghodsi, and Mohamed S. Kamel. Dictionary learning in texture classification. In *Proceedings of the* 8^{th} *international conference on Image analysis and recognition - Volume Part I*, pages 335–343, Berlin, Heidelberg, 2011. Springer-Verlag.

[65] Mehrdad J. Gangeh, Ali Ghodsi, and Mohamed S. Kamel. Kernelized supervised dictionary learning. *IEEE Transactions on Signal Processing*, 61(19):4753–4767, Oct. 2013.

[66] Mehrdad J. Gangeh, Ali Sadeghi-Naini, Mohamed S. Kamel, and Gregory Czarnota. Assessment of cancer therapy effects using texton-based characterization of quantitative ultrasound parametric images. In *Proceedings of the International Symposium on Biomedical Imaging: From Nano to Macro (ISBI)*, pages 1360–1363, 2013.

[67] James Gibson, Athanasios Katsamanis, Matthew P Black, and Shrikanth S Narayanan. Automatic identification of salient acoustic instances in couples' behavioral interactions using diverse density support vector machines. In *INTERSPEECH*, pages 1561–164, 2011.

[68] Michael Glodek, Martin Schels, Günther Palm, and Friedhelm Schwenker. Multiple classifier combination using reject options and markov fusion networks. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 465–472. ACM, 2012.

[69] Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, and Friedhelm Schwenker. Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 359–368. Springer Berlin / Heidelberg, 2011.

[70] Gábor Gosztolya, Róbert Busa-Fekete, and László Tóth. Detecting autism, emotions and social signals using adaboost. In *INTERSPEECH*, pages 220–224, 2013.

[71] Gàbor Gosztolya, Tamás Gròsz, Róbert Busa-Fekete, and László Tóth. Detecting the intensity of cognitive and physical load using adaboost and deep rectifier neural networks. In *INTERSPEECH*, 2014.

[72] A. Gretton, O. Bousquet, A.J. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the* 16^{th} *international conference on Algorithmic Learning Theory (ALT)*, pages 63–77, 2005.

[73] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory*, volume 3734 of *Lecture Notes in Computer Science*, pages 63–77. Springer Berlin / Heidelberg, 2005.

[74] Félix Grèzes, Justin Richards, and Andrew Rosenberg. Let me finish: automatic conflict detection using speaker overlap. In *INTERSPEECH*, pages 200–204, 2013.

[75] M. Grimm, K. Kroschel, and S. Narayanan. The vera am mittag german audio-visual emotional speech database. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 865 –868, 2008.

[76] M. Grimm, E. Mower, K. Kroschel, and S. Narayanan. Combining categorical and primitives-based emotion recognition. In *14th European Signal Processing Conference (EUSIPCO 2006)*, September 2006.

[77] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787 – 800, 2007.

[78] Hatice Gunes, Mihalis A. Nicolaou, and Maja Pantic. Continuous analysis of affect from voice and face. In Albert Ali Salah and Theo Gevers, editors, *Computer Analysis of Human Behavior*, pages 255–291. Springer London, 2011.

[79] Rahul Gupta, Kartik Audhkhasi, Sungbok Lee, and Shrikanth Narayanan. Paralinguistic event detection from speech using probabilistic time-series smoothing and masking. In *INTERSPEECH*, pages 173–177, 2013.

[80] Mark A. Hall. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 359–366, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[81] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2 edition, 2008.

[82] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. *IN NIPS*, 17, 2005.

[83] Matthias Hein and Olivier Bousquet. Kernels, associated structures and generalizations. Technical Report 127, Max Planck Institute for Biological Cybernetics, 2004.

[84] A. Hoerl and R. Kennard. Ridge regression. In *Encyclopedia of Statistical Sciences*, volume 8, pages 129–136. New York: Wiley, 1988.

[85] Dong-Yan Huang, Yongwei Zhu, Dajun Wu, and Rongshan Yu. Detecting intelligibility by linear dimensionality reduction and normalized voice quality hierarchical features. In *INTERSPEECH*, 2012.

[86] Ke Huang and Selin Aviyente. Sparse representation for signal classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 609–616, 2007.

[87] Mark Huckvale. Prediction of cognitive load from speech with the voqal voice quality toolbox for the interspeech 2014 computational paralinguistics challenge. In *INTERSPEECH*, 2014.

[88] Alexei Ivanov and Xin Chen. Modulation spectrum analysis for speaker personality trait recognition. In *INTERSPEECH*, 2012.

[89] Alexei V Ivanov, Giuseppe Riccardi, Adam J Sporka, and Jakub Franc. Recognition of personality traits from human spoken conversations. In *INTERSPEECH*, pages 1549–1552, 2011.

[90] Artur Janicki. Non-linguistic vocalisation recognition based on hybrid gmm-svm approach. In *INTERSPEECH*, pages 153–157, 2013.

[91] How Jing, Ting-Yao Hu, Hung-Shin Lee, Wei-Chen Chen, Chi-Chun Lee, Yu Tsao, and Hsin-Min Wang. Ensemble of machine learning algorithms for cognitive and physical speaker load detection. In *INTERSPEECH*, 2014.

[92] Ittipan Kanluan, Michael Grimm, and Kristian Kroschel. Audio-visual emotion recognition using an emotion space concept. *Signal Processing*, 2008.

[93] Heysem Kaya, Tuğçe Özkaptan, Albert Ali Salah, and Sadık Fikret Gürgen. Canonical correlation analysis and local fisher discriminant analysis based multi-view acoustic feature reduction for physical load prediction. In *INTERSPEECH*, 2014.

[94] Roland Kehrein. The prosody of authentic emotion. In *Speech Prosody 2002*, 2002.

[95] Dacher Keltner and Paul Ekman. Introduction: Expression of emotion. In *Handbook of Affective Sciences*, Series in Affective Science, pages 411–414. Oxford University Press, 2003.

[96] Jangwon Kim, Naveen Kumar, Andreas Tsiartas, Ming Li, and Shrikanth S Narayanan. Intelligibility classification of pathological speech using fusion of multiple subsystems. In *INTERSPEECH*, 2012.

[97] Jonathan Kim, Hrishikesh Rao, and Mark Clements. Investigating the use of formant based features for detection of affective dimensions in speech. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 369–377. Springer Berlin / Heidelberg, 2011.

[98] Samuel Kim, Maurizio Filippone, Fabio Valente, and Alessandro Vinciarelli. Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and gaussian processes. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 793–796. ACM, 2012.

[99] Katrin Kirchhoff, Yuzong Liu, and Jeff Bilmes. Classification of developmental disorders from speech signals using submodular feature selection. In *INTERSPEECH*, pages 187–190, 2013.

[100] Kazuki Kitahara, Shinzi Michiwiki, Miku Sato, Shoichi Matsunaga, Masaru Yamashita, and Kazuyuki Shinohara. Emotion classification of infants' cries using duration ratios of acoustic segments. In *INTERSPEECH*, pages 1573–1576, 2011.

[101] Marcel Kockmann, Lukáš Burget, and Jan Černockỳ. Brno university of technology system for interspeech 2010 paralinguistic challenge. In *INTERSPEECH*, 2010.

[102] Marcel Kockmann, Lukàš Burget, and Jan Černockỳ. Brno university of technology system for interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[103] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *14th Int. Joint Conf. on AI*, volume 2, pages 1137–1143, 1995.

[104] Teun F Krikke and Khiet P Truong. Detection of nonverbal vocalizations using gaussian mixture models: looking for fillers and laughter in conversational speech. In *INTERSPEECH*. International Speech Communication Association, 2013.

[105] Jia Min Karen Kua, Vidhyasaharan Sethu, Phu Le, and Eliathamby Ambikairajah. The unsw submission to interspeech 2014 compare cognitive load challenge. In *INTERSPEECH*, 2014.

[106] Solomon Kullback. *Information Theory and Statistics*. John Wiley & Sons, 1959.

[107] Petri Laukka, Daniel Neiberg, Mimmi Forsell, Inger Karlsson, and Kjell Elenius. Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language*, 25(1):84 – 104, 2011.

[108] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1294–1309, July 2009.

[109] Chi-Chun Lee, Matthew Black, Athanasios Katsamanis, Adam C Lammert, Brian R Baucom, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *INTERSPEECH*, pages 793–796, 2010.

[110] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition using a hierarchical binary decision tree approach. In *INTERSPEECH*, volume 53, pages 1162–1171. Elsevier, 2009.

[111] Hung-yi Lee, Ting-yao Hu, How Jing, Yun-Fan Chang, Yu Tsao, Yu-Cheng Kao, and Tsang-Long Pao. Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition. In *INTERSPEECH*, pages 215–219, 2013.

[112] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. Multimedia content processing through cross-modal association. In *Proceedings of the* 11^{th} *ACM international conference on Multimedia*, pages 604–611, 2003.

[113] Ming Li. Automatic recognition of speaker physical load using posterior probability based features from acoustic and phonetic tokens. In *INTERSPEECH*, 2014.

[114] Ming Li, Chi-Sang Jung, and Kyu Jeong Han. Combining five acoustic level modeling methods for automatic speaker age and gender recognition. In *INTERSPEECH*, pages 2826–2829, 2010.

[115] Florian Lingenfelser, Johannes Wagner, Thurid Vogt, Jonghwa Kim, and Elisabeth André. Age and gender classification from speech using decision level fusion and ensemble based techniques. In *INTERSPEECH*, volume 10, pages 2798–2801, 2010.

[116] Dingchao Lu and Fei Sha. Predicting likability of speakers with gaussian processes. In *INTERSPEECH*, 2012.

[117] Iker Luengo, Eva Navas, and Inmaculada Hernáez. Combining spectral and prosodic information for emotion recognition in the interspeech 2009 emotion challenge. In *INTERSPEECH*, pages 332–335, 2009.

[118] H. Lütkepohl. *Handbook of Matrices*. John Wiley & Sons, 1996.

[119] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1033–1040, 2008.

[120] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, Jan. 2008.

[121] S. Mallat. *A Wavelet Tour of signal Processing: The Sparse Way*. Academic Press, third edition, 2009.

[122] David Martınez, Dayana Ribas, Eduardo Lleida, Alfonso Ortega, and Antonio Miguel. Suprasegmental information modelling for autism disorder spectrum and specific language impairment classification. In *INTERSPEECH*, 2013.

[123] G. McKeown, M.F. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079 –1084, july 2010.

[124] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5 –17, jan.-march 2012.

[125] A. Mehrabian. *Basic dimensions for a general psychological theory: implications for personality, social, environmental, and developmental studies*. Social Environmental and Developmental Studies. Oelgeschlager, Gunn & Hain, 1980.

[126] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14:261–292, 1996.

[127] Hugo Meinedo and Isabel Trancoso. Age and gender classification using fusion of acoustic and prosodic features. In *INTERSPEECH*, pages 2818–2821, 2010.

[128] Hongying Meng and Nadia Bianchi-Berthouze. Naturalistic affective expression classification by a multi-stage approach based on hidden markov models. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 378–387. Springer Berlin / Heidelberg, 2011.

[129] Gelareh Mohammadi, Alessandro Vinciarelli, and Marcello Mortillaro. The voice of personality: Mapping nonverbal vocal behavior into trait attributions. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 17–20. ACM, 2010.

[130] Claude Montacié and Marie-José Caraty. Pitch and intonation contribution to speakers' traits classification. In *INTERSPEECH*, 2012.

[131] Claude Montacié and Marie-José Caraty. High-level speech event analysis for cognitive load classification. In *INTERSPEECH*, 2014.

[132] Frank Moosmann, Bill Triggs, and Frédéric Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Advances in Neural Information Processing Systems (NIPS)*, pages 985–992, 2006.

[133] Donn Morrison, Ruili Wang, and Liyanage C. De Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98 – 112, 2007.

[134] Emily Mower, Kyu Jeong Han, Sungbok Lee, and Shrikanth S Narayanan. A cluster-profile representation of emotion using agglomerative hierarchical clustering. In *INTERSPEECH*, pages 797–800, 2010.

[135] Daniel Neiberg, Kjell Elenius, and Kornel Laskowski. Emotion recognition in spontaneous speech using gmms. In *INTERSPEECH-2006*, 2006.

[136] Daniel Neiberg and Joakim Gustafson. Predicting speaker changes and listener responses with and without eye-contact. In *INTERSPEECH*, pages 1565–1568, 2011.

[137] Daniel Neiberg, Petri Laukka, and Hillary Anger Elfenbein. Intra-, inter-, and cross-cultural classification of vocal affect. In *INTERSPEECH*, pages 1581–1584, 2011.

[138] Phuoc Nguyen, Trung Le, Dat Tran, Xu Huang, and Dharmendra Sharma. Fuzzy support vector machines for age and gender classification. In *INTERSPEECH*, pages 2806–2809, 2010.

[139] M.A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on*, 2(2):92 –105, april-june 2011.

[140] Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 501–508. ACM, 2012.

[141] Narichika Nomoto, Masafumi Tamoto, Hirokazu Masataki, Osamu Yoshioka, and Satoshi Takahashi. Anger recognition in spoken dialog using linguistic and para-linguistic information. In *INTERSPEECH*, pages 1545–1548, 2011.

[142] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41(4):603 – 623, 2003.

[143] Tin Lay Nwe, Nguyen Trung Hieu, and Bin Ma. On the use of bhattacharyya based gmm distance and neural net features for identification of cognitive load levels. In *INTERSPEECH*, 2014.

[144] Catharine Oertel, Stefan Scherer, and Nick Campbell. On the use of multimodal cues for the prediction of involvement in spontaneous conversation. In *INTERSPEECH*, 2011.

[145] Jieun Oh, Eunjoon Cho, and Malcolm Slaney. Characteristic contours of syllabic-level units in laughter. In *INTERSPEECH*, pages 158–162, 2013.

[146] Derya Ozkan, Stefan Scherer, and Louis-Philippe Morency. Step-wise emotion recognition using concatenated-hmm. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 477–484. ACM, 2012.

[147] Shifeng Pan, Jianhua Tao, and Ya Li. The casia audio emotion recognition method for audio/visual emotion challenge 2011. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 388–395. Springer Berlin / Heidelberg, 2011.

[148] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238, 2005.

[149] Humberto Pérez-Espinosa, Carlos A. Reyes-Garcia, and Luis Villase nor Pineda. Acoustic feature selection and classification of emotions in speech using a 3d continuous emotion model. *Biomedical Signal Processing and Control*, 7(1):79 – 87, 2012.

[150] R. W. Picard. Affective Computing for HCI. *Human-Computer Interaction: Ergonomics and User Interfaces*, 1:829–833, 1999.

[151] Rosalind W. Picard. Affective computing. Technical Report 321, M.I.T Media Laboratory Perceptual Computing Section, 1995.

[152] Oudeyer Pierre-Yves. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2):157 – 183, 2003.

[153] Santiago Planet, Ignasi Iriondo Sanz, Joan Claudi Socoró, Carlos Monzo, and Jordi Adell. Gtm-url contribution to the interspeech 2009 emotion challenge. In *INTERSPEECH*, pages 316–319, 2009.

[154] Jouni Pohjalainen and Paavo Alku. Filtering and subspace selection for spectral features in detecting speech under physical stress. In *INTERSPEECH*, 2014.

[155] Jouni Pohjalainen, Serdar Kadioglu, and Okko Räsänen. Feature selection for speaker traits. In *INTERSPEECH*, 2012.

[156] Tim Polzehl, Katrin Schoenenberg, Sebastian Moller, Florian Metze, Gelareh Mohammadi, and Alessandro Vinciarelli. On speaker-independent personality perception and prediction from speech. In *INTERSPEECH*, 2012.

[157] Tim Polzehl, Shiva Sundaram, Hamed Ketabdar, Michael Wagner, and Florian Metze. Emotion classification in children's speech using fusion of acoustic and linguistic features. In *INTERSPEECH*, pages 340–343, 2009.

[158] Royi Porat, Dan Lange, and Yaniv Zigel. Age recognition based on speech signals using weights supervector. In *INTERSPEECH*, pages 2814–2817, 2010.

[159] SR Mahadeva Prasanna and D Govind. Analysis of excitation source information in emotional speech. In *INTERSPEECH*, pages 781–784, 2010.

[160] S. Ramakrishnan and Ibrahiem El Emary. Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, pages 1–12, 2011.

[161] Geovany Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 396–406. Springer Berlin / Heidelberg, 2011.

[162] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3508, 2010.

[163] S.R. Rao, R. Tron, R. Vidal, and Yi Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[164] Okko Räsänen and Jouni Pohjalainen. Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech. In *INTERSPEECH*, pages 210–214, 2013.

[165] J.D. Rodriguez, A. Perez, and J.A. Lozano. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(3):569–575, 2010.

[166] Jia Rong, Gang Li, and Yi-Ping Phoebe Chen. Acoustic feature selection for automatic emotion recognition from speech. *Information Processing & Management*, 45(3):315 – 328, 2009.

[167] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[168] S. Saitō and K. Nakata. *Fundamentals of speech signal processing*. Academic Press, 1985.

[169] Michelle Hewlett Sanchez, Aaron Lawson, Dimitra Vergyri, and Harry Bratt. Multi-system fusion of extended context prosodic and cepstral features for paralinguistic speaker trait classification. In *INTERSPEECH*, 2012.

[170] Arman Savran, Houwei Cao, Miraj Shah, Ani Nenkova, and Ragini Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 485–492. ACM, 2012.

[171] Aya Sayedelahl, Pouria Fewzee, Mohamed Kamel, and Fakhri Karray. Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 407–414. Springer Berlin / Heidelberg, 2011.

[172] Klaus R. Scherer, Tom Johnstone, and Gundrun Klasmeyer. Vocal expression of emotion. In *Handbook of Affective Sciences*, Series in Affective Science, pages 433–456. Oxford University Press, 2003.

[173] B. Schuller. Recognizing affect from linguistic information in 3d continuous space. *Affective Computing, IEEE Transactions on*, 2(4):192 –205, oct.-dec. 2011.

[174] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll. Speaker independent speech emotion recognition by ensemble classification. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 864 –867, july 2005.

[175] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 1, pages 577–580, may 2004.

[176] B. Schuller, 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. In *INTERSPEECH*, 2013.

[177] B. Schuller, B. Vlasenko, D. Arsic, G. Rigoll, and A. Wendemuth. Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 1333 –1336, 2008.

[178] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. Acoustic emotion recognition: A benchmark comparison of performances. In *Automatic Speech Recognition Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 552 –557, 2009.

[179] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *Affective Computing, IEEE Transactions on*, 1(2):119 –131, 2010.

[180] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll. Using multiple databases for training in emotion recognition: To unite or to vote? In *INTERSPEECH*, 2011.

[181] Björn Schuller and Laurence Devillers. Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm. In *INTERSPEECH*, pages 801–804, 2010.

[182] Bjorn Schuller, Ronald Muller, Florian Eyben, Jurgen Gast, Benedikt Hornler, Martin Wollmer, Gerhard Rigoll, Anja Hothker, and Hitoshi Konosu. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12):1760 – 1774, 2009.

[183] Björn Schuller, Ronald Müller, Manfred Lang, and Gerhard Rigoll. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *INTERSPEECH-2005*, 2005.

[184] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315. Citeseer, 2009.

[185] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, and Shrikanth S Narayanan. The interspeech 2010 paralinguistic challenge. In *INTERSPEECH*, pages 2794–2797, 2010.

[186] Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load. In *INTERSPEECH*, volume 2014, 2014.

[187] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss. The interspeech 2012 paralinguistic challenge. In *INTERSPEECH*, pages 254–257, 2012.

[188] Björn Schuller, Michel Valstar, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012 – the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 449–456. ACM, 2012.

[189] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 415–424. Springer Berlin / Heidelberg, 2011.

[190] Vidhyasaharan Sethu, Julien Epps, Eliathamby Ambikairajah, and Haizhou Li. Gmm based speaker variability compensated system for interspeech 2013 compare emotion challenge. In *INTERSPEECH*, pages 205–209, 2013.

[191] Krzysztof Ślot, Jaroslaw Cichosz, and Lukasz Bronakowski. Emotion recognition with poincare mapping of voiced-speech segments of utterances. In *Artificial Intelligence and Soft Computing – ICAISC 2008*, volume 5097 of *Lecture Notes in Computer Science*, pages 886–895. Springer Berlin / Heidelberg, 2008.

[192] Ian Sneddon, Gary McKeown, Margaret McRorie, and Tijana Vukicevic. Cross-cultural patterns in dynamic ratings of positive and negative natural emotional behaviour. *PLoS ONE*, 6(2), 02 2011.

[193] Catherine Soladié, Hanan Salam, Catherine Pelachaud, Nicolas Stoiber, and Renaud Séguier. A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 493–500. ACM, 2012.

[194] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *Affective Computing, IEEE Transactions on*, 3(1):42 –55, 2012.

[195] Lin Song, Peter Langfelder, and Steve Horvath. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC bioinformatics*, 14(1):5, 2013.

[196] L. Sørensen, M. J. Gangeh, S. B. Shaker, and M. de Bruijne. Texture classification in pulmonary CT. In A. El-Baz and J. S. Sure, editors, *Lung Imaging and Computer Aided Diagnosis*, pages 343–367. CRC Press, 2007.

[197] Anthony P Stark, Alireza Bayestehtashk, Meysam Asgari, and Izhak Shafran. Interspeech pathology challenge: Investigations into speaker and sentence specific effects. In *INTERSPEECH*, 2012.

[198] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5688 –5691, may 2011.

[199] Rui Sun and Elliot Moore. Investigating glottal parameters and teager energy operators in emotion recognition. In Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, editors, *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 425–434. Springer Berlin / Heidelberg, 2011.

[200] A. Tarasov and S.J. Delany. Benchmarking classification models for emotion recognition in natural speech: A multi-corporal study. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 841 –846, march 2011.

[201] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[202] Rob Tibshirani, Trevor Hastie, and Jerome H. Friedman. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(01), 2010.

[203] K. P. Truong, D. A. van Leeuwen, M. A. Neerincx, and F. M. G. de Jong. Arousal and valence prediction in spontaneous emotional speech: Felt versus perceived emotion. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009), Brighton, United Kingdom*, pages 2027–2030. International Speech Communication Association, 2009.

[204] Laurens van der Maaten. Audio-visual emotion challenge 2012: a simple approach. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 473–476, New York, NY, USA, 2012. ACM.

[205] Maarten Van Segbroeck, Ruchir Travadi, Colin Vaz, Jangwon Kim, Matthew P Black, Alexandros Potamianos, and Shrikanth S Narayanan. Classification of cognitive load from speech using an i-vector framework. In *INTERSPEECH*, 2014.

[206] H. D. Vankayalapati, K. R. Anne, and K. Kyamakya. Decision level fusion of visual and acoustic features of the driver for real-time driver monitoring system. In *Transactions on Computers and Intelligent Systems*, volume 3, pages 14–22. International Society for Advanced Science and Technology (ISAST), July 2011.

[207] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis*, 62(1-2):61–81, 2005.

[208] M. Varma and A. Zisserman. A statistical approach to material classification using image patch exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2032–2047, Nov. 2009.

[209] Dimitrios Ververidis and Constantine Kotropoulos. Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition. *Signal Processing*, 88(12):2956 – 2970, 2008.

[210] Bogdan Vlasenko, Dmytro Prylipko, David Philippou-Hübner, and Andreas Wendemuth. Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. In *INTERSPEECH*, pages 1577–1580, 2011.

[211] Thurid Vogt and Elisabeth André. Exploring the benefits of discretization of acoustic features for speech emotion recognition. In *INTERSPEECH*, pages 328–331, 2009.

[212] Johannes Wagner, Florian Lingenfelser, and Elisabeth André. A frame pruning approach for paralinguistic recognition tasks. In *INTERSPEECH*, 2012.

[213] Johannes Wagner, Florian Lingenfelser, and Elisabeth André. Using phonetic patterns for detecting social cues in natural conversations. In *INTERSPEECH*, pages 168–172, 2013.

[214] Yongjin Wang, Ling Guan, and A.N. Venetsanopoulos. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia*, 14(3):597–607, June 2012.

[215] Benjamin Weiss and Felix Burkhardt. Is' not bad'good enough? aspects of unknown voices' likability. In *INTERSPEECH*, 2012.

[216] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In 10^{th} *IEEE International Conference on Computer Vision (ICCV)*, pages 1800–1807, 2005.

[217] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie. Data-driven clustering in emotional space for affect recognition using discriminatively trained lstm networks. In *INTERSPEECH-2009*, 2009.

[218] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *Selected Topics in Signal Processing, IEEE Journal of*, 4(5):867 –881, oct. 2010.

[219] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH-2008*, pages 597–600, 2008.

[220] Martin Wöllmer, Moritz Kaiser, Florian Eyben, Björn Schuller, and Gerhard Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 2012.

[221] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, Feb. 2009.

[222] Chung-Hsien Wu and Wei-Bin Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *Affective Computing, IEEE Transactions on*, 2(1):10 –21, 2011.

[223] D. Wu, T. D. Parsons, and S. S. Narayanan. Acoustic feature analysis in speech emotion primitives estimation. In *INTERSPEECH*, 2010.

[224] Dongrui Wu. Genetic algorithm based feature selection for speaker trait classification. In *INTERSPEECH*, 2012.

[225] Dongrui Wu, T.D. Parsons, E. Mower, and S. Narayanan. Speech emotion estimation in 3d space. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 737 –742, july 2010.

[226] Siqing Wu, T.H. Falk, and Wai-Yip Chan. Automatic recognition of speech emotion using long-term spectro-temporal features. In *Digital Signal Processing, 2009 16th International Conference on*, pages 1 –6, july 2009.

[227] Siqing Wu, Tiago H. Falk, and Wai-Yip Chan. Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53(5):768 – 785, 2011.

[228] J. Xie, L. Zhang, J. You, and D. Zhang. Texture classification via patch-based sparse texton learning. In 17th *IEEE International Conference on Image Processing (ICIP)*, pages 2737–2740, 2010.

[229] Sherif Yacoub, Steve Simske, Xiaofan Lin, and John Burns. Recognition of emotions in interactive voice response systems. In *EUROSPEECH-2003*, pages 729–732, 2003.

[230] Jianchao Yang, J. Wright, T.S. Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, Nov. 2010.

[231] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In 13th *IEEE International Conference on Computer Vision (ICCV)*, pages 543–550, 2011.

[232] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. Metaface learning for sparse representation based face recognition. In 17th *IEEE International Conference on Image Processing (ICIP)*, pages 1601–1604, 2010.

[233] Lan-Ying Yeh and Tai-Shih Chi. Spectro-temporal modulations for robust speech emotion recognition. In *INTERSPEECH*, pages 789–792, 2010.

[234] C. Yu, P. M. Aoki, and A. Woodruff. Detecting user engagement in everyday conversations. In *Int'l Conf. on Spoken Language Processing (ICSLP)*, volume 2, pages 1329–1332, October 2004.

[235] Sungrack Yun and C.D. Yoo. Loss-scaled large-margin gaussian mixture models for speech emotion classification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):585 –598, 2012.

[236] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1601–1608, 2004.

[237] Qiang Zhang and Baoxin Li. Discriminative K-SVD for dictionary learning in face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2691–2698, 2010.

[238] Yin Zhang and Zhi-Hua Zhou. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):14, 2010.

[239] Zixing Zhang, F. Weninger, M. Wollmer, and B. Schuller. Unsupervised learning in cross-corpus acoustic emotion recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 523 –528, dec. 2011.

[240] Cheng Zhong, Zhenan Sun, and Tieniu Tan. Robust 3D face recognition using learned visual codebook. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 2007.

[241] Xinhui Zhou, Daniel Garcia-Romero, Nima Mesgarani, Maureen C Stone, Carol Y Espy-Wilson, and Shihab A Shamma. Automatic intelligibility assessment of pathologic speech in head and neck cancer based on auditory-inspired spectro-temporal modulations. In *INTERSPEECH*, 2012.

[242] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.