Automatic Speech Emotion Recognition- Feature space Dimensionality and Classification Challenges

By

Abdulbasit Kamil Faeq Al-Talabani

Department of Applied Computing

University of Buckingham

United Kingdom

A thesis

submitted for the degree of Doctor of philosophy in Applied Computing to the school of science in the University of Buckingham.

September/2015

Abstract

In the last decade, research in Speech Emotion Recognition (SER) has become a major endeavour in Human Computer Interaction (HCI), and speech processing. Accurate SER is essential for many applications, like assessing customer satisfaction with quality of services, and detecting/assessing emotional state of children in care. The large number of studies published on SER reflects the demand for its use. The main concern of this thesis is the investigation of SER from a pattern recognition and machine learning points of view. In particular, we aim to identify appropriate mathematical models of SER and examine the process of designing automatic emotion recognition schemes. There are major challenges to automatic SER including ambiguity about the list/definition of emotions, the lack of agreement on a manageable set of uncorrelated speech-based emotion relevant features, and the difficulty of collected emotion-related datasets under natural circumstances. We shall initiate our work by dealing with the identification of appropriate sets of emotion-related features/attributes extractible from speech signals as considered from psychological and computational points of views. We shall investigate the use of pattern-recognition approaches to remove redundancies and achieve compactification of digital representation of the extracted data with minimal loss of information. The thesis will include the design of new or complement existing SER schemes and conduct large sets of experiments to empirically test their performances on different databases, identify advantages, and shortcomings of using speech alone for emotion recognition.

Existing SER studies seem to deal with the ambiguity/dis-agreement on a "limited" number of emotion-related features by expanding the list from the same speech signal source/sites and apply various feature selection procedures as a mean of reducing redundancies. Attempts are made to discover more relevant features to emotion from speech. One of our investigations focuses on proposing a newly sets of features for SER, extracted from Linear Predictive (LP)-residual speech. We shall demonstrate the usefulness of the proposed relatively small set of features by testing the performance of an SER scheme that is based on fusing our set of features with the existing set of thousands of features using common machine learning schemes of Support Vector Machine (SVM) and Artificial Neural Network (ANN).

The challenge of growing dimensionality of SER feature space and its impact on increased model complexity is another major focus of our research project. By studying the pros and cons of the commonly used feature selection approaches, we argued in favour of meta-feature selection and developed various methods in this direction, not only to reduce dimension, but also to adapt and de-correlate emotional feature spaces for improved SER model recognition accuracy. We used Principal Component Analysis (PCA) and proposed Data Independent PCA (DIPCA) by training on independent emotional and non-emotional datasets. The DIPCA projections, especially when extracted from speech data coloured with different emotions or from Neutral speech data, had comparable capability to the PCA in terms of SER performance. Another adopted approach in this thesis for dimension reduction is the Random Projection (RP) matrices, independent of training data. We have shown that some versions of RP with SVM classifier can offer an adaptation space for Speaker Independent SER that avoid over-fitting and hence improves recognition accuracy. Using PCA trained on a set of data, while testing on emotional data features, has significant implication for machine learning in general.

The thesis other major contribution focuses on the classification aspects of SER. We investigate the drawbacks of the well-known SVM classifier when applied to a pre-processed data by PCA and RP. We shall demonstrate the advantages of using the Linear Discriminant Classifier (LDC) instead especially for PCA de-correlated meta-features. We initiated a variety of LDC-based ensembles classification, to test performance of scheme using a new form of bagging different subsets of meta-feature subsets extracted by PCA with encouraging results.

The experiments conducted were applied on two benchmark datasets (Emo-Berlin and FAU-Aibo), and an in-house dataset in the Kurdish language. Recognition accuracy achieved by are significantly higher than the state of art results on all datasets. The results, however, revealed a difficult challenge in the form of persisting wide gap in accuracy over different datasets, which cannot be explained entirely by the differences between the natures of the datasets. We conducted various pilot studies that were based on various visualizations of the confusion matrices for the “difficult” databases to build multi-level SER schemes. These studies provide initial evidences to the presence of more than one “emotion” in the same portion of speech. A possible solution may be through presenting recognition accuracy in a score-based

measurement like the spider chart. Such an approach may also reveal the presence of Doddington zoo phenomena in SER.

Acknowledgement

I would like to express my sincere gratitude to my supervisors Prof. Sabah Jassim and Dr. Harin Sellahewa for the continuous support of my Ph.D study, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis.

My sincere thanks also goes to the whole staff and researcher students at the University of Buckingham who provided me an opportunity to join many valuable stimulating discussions.

I would like to thank my family: my wife and to my sister and brothers for supporting me spiritually throughout writing this thesis and my life in general.

Contents

List of Abbreviation

Abbreviation	Description
AERR	Average Emotion Recognition Rate
ANN	Artificial Neural Network
AUC	Area Under Curve
BRP	Binary Random Projection
CFS	Correlation based Feature Selection
CM	Confusion Matrix
CS	Compressive Sensing
DAG	Directed Acyclic Graph
DEA	De-noising Auto-encoder
DIPCA	Data Independent Principal Component Analysis
DrSVM	Doubly regularized SVM
EEG	electroencephalogram
EER	Equal Error Rates
ES	Excitation Source features
F0	Fundamental frequency
FT	Fourier transform
GR	Gaussian Random
HCI	Human Computer Interaction
HOG	Histogram of Oriented Gradients
i.i.d.	Independent and identically distributed
ISVM	Iterative SVM
IW	Importance Weight
k-NN	k-Nearest Neighbour
l_1-SVM	l_1 norm within SVM
LDA	Linear Discriminant Analysis
LDC	Linear Discriminant Classifier
LLD	Low Level Descriptors
LOO	Leave One Out
LOSO	Leave One Speaker Out
LP	Linear Predictive
LPCC	Linear Predictive Cypstrum Coefficient
MC	Multi-level Classifier
MER	Music Emotion Recognition
MFCC	Mel Frequency Cypstrum Coefficient
MOS	Mean Opinion Score
MSE	Mean Square Error
NDMS	Non-metric Multi-Dimensional Scaling
NP	No-deterministic Polynomial-time
OP	6552 features extracted using OpenEAR software

OVA	One Versus All
OVO	One Versus One
PCA	Principal Component Analysis
PVQ	P models Versus Q models
RBF	Radial Basis Function
RFS	Random Feature Selection
RIP	Restricted Isometric Property
ROC	Receiver Operating Characteristic
RP	Random Projection
RRIP	Random Restricted Isometric Projection
SBS	Sequential Backward Selection
SD	Speaker Dependent
SER	Speech Emotion Recognition
SFFS	Sequential Floating Feature Selection
SFS	Sequential Forward Selection
SI	Speaker Independent
SMO	Sequential Minimal Optimization
SMOTE	Synthetic Minority Over-sampling Technique
SN	Speaker Normalisation
SVM	Support Vector Machine
T_n	Toeplitz
UAR	Un-weighted Average Recall
UDT	Unbalanced Decision Tree
WAR	Weighted Average Recall
WOCOR	Wavelet Octave Coefficients Of Residual

List of Figures

List of Tables

Chapter One
Introduction

Speech is an important way for people to communicate with one another, and emotion in speech can change the meaning of sentences as perceived by others, and the intentions of the speaker. Spoken text can have several different meanings, depending on how it is said. For example, with the word "really" in English, a speaker can ask a question, express either admiration or disbelief, or make a definitive statement. An understanding of text alone cannot always successfully interpret the meaning of a spoken utterance. Emotion modulates the choice of words and the tone of voice in speech as well as in many other human modes of interaction with other humans and/or computers. Designing automatic systems that have the ability of recognize emotion automatically is deemed to be useful for many applications including in healthcare settings and Human Computer Interaction (HCI).

Speech Emotion Recognition (SER) is no longer a side issue. In the last decade, research in SER has become a major endeavour in HCI and in speech processing. The large number of studies published with regard to SER reflects the demand for its use (see Figure 1-1). The main concern of this thesis is the investigation of the ability to recognise emotions within speech from a pattern recognition point of view. In particular, the aim is to design automatic emotion recognition schemes, test their performance in terms of different databases, and identify the advantages and shortcomings of using speech alone for emotion recognition. Our investigations focuses mainly on emotion-related feature extraction from speech, and the use of different classifiers and their fusion. One ultimate objective is to determine the contribution of speech to the emotional state of speakers.

In section (1.1) of this chapter we will focus on the some application regarding SER. The chapter will also present the challenges faced by SER as a pattern recognition task in sections (1.2) and (1.3). The contribution of the thesis is presented in section (1.4) followed by the publications related to this thesis in section (1.5), and finally the organisation of the thesis is shown in section (1.6).

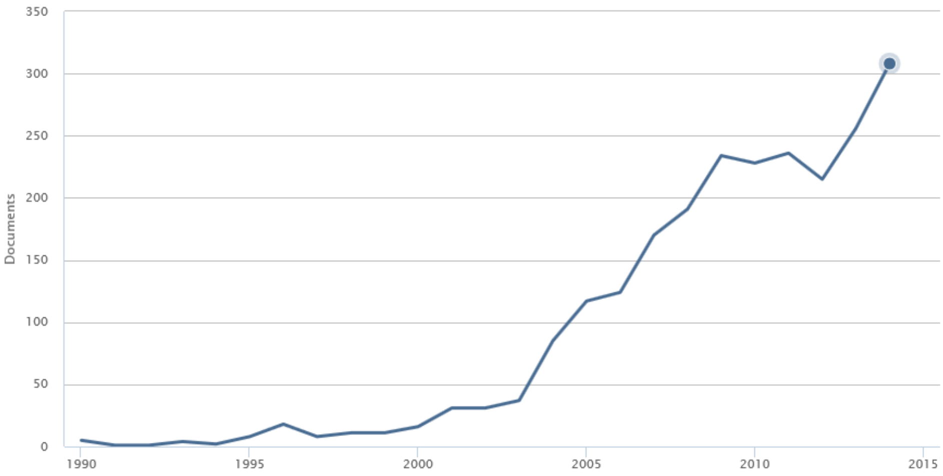

Figure 1-1: Number of publications on SER from 1990-2014

1.1 Speech based emotion recognition applications

Studying the emotional state of speakers from their speech signals has emerged as one of the most important areas of speech research. Any speech system can be made more effective by incorporating emotion processing/analysis (Koolagudi & Rao, 2012b). Knowledge of the emotional state of a speaker helps the listener/system understand the meaning of a message. For example, speech recognition systems that are designed to assess stressed-speech in aircraft cockpits have been found to achieve better performance than traditional emotion-ignoring speech recognition systems (Hansen & Cairns, 1995).

We recognize that speech is only one of many modes of interaction with regard to human beings interacting with one another and with machines that are modulated by emotional states. Moreover the expression of emotion is influenced by many factors including culture, personal experience, and by mental health state. Consequently, one cannot expect SER to be other than of limited success in most cases. However, in many applications, speech is the only available mode of interaction, and consequently modestly accurate SER can benefit such applications. For example, customers care interaction systems that assess customer’s satisfaction and quality of service. It is also helpful in fostering interaction systems, which aim to detect a child’s emotional state (Lee , et al., 2011). Another, possible application of SER is hostage negotiation standoffs where speech is most likely to be the only available mode of communication. Building a dialogue system for natural speech is becoming an interesting area of research that could benefit from SER. Most existing applications with regard to automatic dialogue systems are reported to be restricted to using binary responses (i.e. 'yes' and 'no'), or designed for a specific area with a limited vocabulary. Consequently, responding in a more "natural" way to different form of emotional speech is still a serious challenge and one for which there is an ongoing demand[1] (Steidl, 2009).

In healthcare, people with anxiety, depression, and stress are frequently asked by psychologists to record their mood changes throughout the day. An application called Xpression, developed by U.K.-based EI Technologies, helps such people by sending a block of 200ms speech voice to the designed server to detect the embedded emotion and then report the changes in their mood to a psychologist every day[2]. SER is also used to help hearing-impaired people to speak naturally by designing a system, which can tell the users about their speech state. This system is able to train the user to improve his/her skills in terms of natural speech (Pao , et al., 2005). Accurate identification of a patient's emotional state is also reported to be necessary for psychotherapists (Paulo , et al., 1999); consequently, it is suggested that an emotion recognition framework be integrated into a speech therapist system in (Schipor , et al., 2011). A content delivery system can also benefit from tracking emotional information contained in the objects (video, music) to improve the delivery quality (Malandrakis , et al., 2011). The effect of emotion recognition and regulation on intercultural adjustment is discussed in a study, which used different groups of students from different cultural environments studying at US universities. The study found that emotional recognition has an independent effect on intercultural adjustment (Yoo, et al., 2006). Such a high number of applications encourage further study to improve the performance of SER.

1.2 Challenges of SER

Raw speech signals captured by a microphone over a period of time are represented by a 1-dimentional waveform consisting of different frequencies, the contributions of which vary over time. Different frequency domain transforms can be used to obtain the frequency domain representation of the signal at multiple scales, but identifying the speaker's emotional state from temporal domain or frequency domain representations is a challenge that has attracted research scientists from different

[1] A comprehensive discussion on some real challenges facing automatic call responding system is presented in (Steidl, 2009) pp. 1-3

[2] http://ipglab.com/2013/03/04/xpression-app-tracks-your-mood/

disciplines. Variation in voice tones, as well as internal physiological changes while uttering a sentence (or even a single word), combine to generate the speaker's emotional state. The speech signal, even when uttering a particular word, is affected by external factors such as gender, age, culture and health. In other words, the speech signal contributes to the speaker's emotional state, but cannot singularly discriminate emotions. The first challenge is to identify emotion-relevant features that can be extracted from the raw signal or from its frequency domain. This challenge remains an active area of research and debate.

Many recent studies have observed that emotion-related information in speech is spread along different kinds of features. This could be due to acoustic variability as a consequence of the existence of different sentences, speakers, speaking styles, and speaking rates (El Ayadi , et al., 2011). Although the number of investigated features has been growing, there are still efforts being made to explore emotion-relevant features from new sites of the speech signal. Chapter four of this thesis focuses on suggesting a new set of features for SER use. The features are extracted from the Linear Predicted (LP)-residual signal.

Researchers have adopted large feature set (thousands), whether directly as in (Hassan & Damper, 2012; Schuller, et al., 2009b), and/or followed by feature selection steps such as are used in (Batliner , et al., 2011). Feature selection applied to high dimensional data is costly and complicated due to the complexity of the optimization that targets a suitable feature subset among a high number of features, especially when using the wrapper methods. Unlike the wrapper feature selection methods, filter based feature selection is not based on classification decision, but on some data characteristics such as correlation or entropy. Filters are reported to be more convenient for high dimensional data (Yu & Liu, 2003). However filter-based feature selection methods are not necessarily suitable for all classifiers, and feature selection cut off points could lead the system to ignore some "important" information included in the non-selected features.

The features can also be selected in a transformed space. Transforming the data in the feature space into another subspace is usually based on a specific transformation map, which aims to help in features de-correlating, samples classifying, or at least reducing dimensions. The new selected features in the transformed space could be referred to as a meta-feature. However, the extraction of the meta-feature is difficult to provide meaning about the more suitable original features, which leads to them being avoided by the studies that target emotional analysis.

Recently, there have been promising efforts in terms of finding various approaches for interpreting meta-features. For instance Simmons et al. (Simmons, et al., 2015) propose a hybrid approach that uses a mutual information-based statistic to have a biologically meaningful interpretation for the Principal Component Analysis (PCA) output. These studies are encouraging further investigation on the use of meta-features for SER. However this thesis is not focussing on meta-feature interpretation, but rather in chapter five the extraction and selection of meta-features is investigated. We shall see how a limited number of the meta-features of emotional speech shows an ability to exploit the information spread in a large set of features.

Beside the challenges related to appropriate SER feature vector representation and extraction, testing the performance of any developed SER scheme faces another serious obstacle. Like any pattern-recognition task, developing and testing the performance of emotion classification schemes needs access to a sufficiently large dataset of feature vectors that are accurately labelled by their emotions or emotion-related states. Creating such databases of speech signals assumes that the collected data do represent real life emotional speech, and can be easily labelled by human experts. This is a tough challenge that has frequently been discussed, questioned, and researched. The traditional method of data collection started with gathering acted emotion data (Batliner , et al., 2011). These data are normally designed by asking actors to repeatedly express different specific emotions, and suppress all other emotions, while uttering the same sentences chosen from a well-selected set of texts. The reliability of the emotional labels is dependent on how good and skilful the actors are. However, the concern is about the extent to which the data is a good representation of real life emotions. For this purpose, trying to collect “natural” emotional data requires repeated recorded speeches of subjects while expressing specific emotions without being aware of the recording. However, this approach might have serious ethical implications. Consequently, researchers have attempted to create non-acted databases by collecting non-prompted data (Steidl, 2009). Capturing the “natural” emotional speech portions and then labelling them in terms of their emotion is challenging. For instance, in designing the FAU-Aibo database, children were convinced that they were controlling a pet robot by their speech instructions, while in reality the robot pet was controlled by a special wizard. Consequently, obeying or disobeying the children resulted in emotional speech states, which were recorded and then labelled to emotional-related states based on the decision of five experts. However such a design still suffers from producing a monologue instead of a dialogue (Schuller, et al., 2011b).

SER accuracy is significantly different from one database to another, due to the design (acted, elicited, or non-acted), the speaker factor (gender and\or age), and the purpose of creating the dataset (analysis, synthesis, recognition). Therefore the recognition performance of a SER system achieves comparable accuracy rate to human raters in subjective listening tests (Vogt , et al., 2008).

Perfect recognition of emotion is not easy even by humans when listening to one another; sometimes the human cannot recognize his own innermost emotion. In fact, some aspects of internal feelings remain hidden and do not appear in the speech, especially when the speaker likes to do that. Therefore, a computer-based system cannot do beyond what is observed from the speech sample input (Picard , et al., 2001).

In this thesis we have tested our developed schemes both on acted and non-acted databases, including an in-house acted database, and we shall attempt to measure and explain the expected discrepancy in accuracy rates achieved by the two types.

1.3 SER as a pattern recognition task

SER follows the same steps of pattern recognition (Figure 1-2), starting with pre-processing the input data by extracting and selecting suitable features, followed by the classification step. The classification techniques have a clear impact on the model's complexity and performance. Some classifiers are well known in improving model performance such as the Support Vector Machine (SVM), but in contrast, the SVM's complexity may be considered as a serious disadvantage for some applications. Some other classifiers are simple in terms of computation and implementation, but may not perform well every time like the k-Nearest Neighbour (k-NN), and Linear Discriminant Classifier (LDC), especially in the case of a high

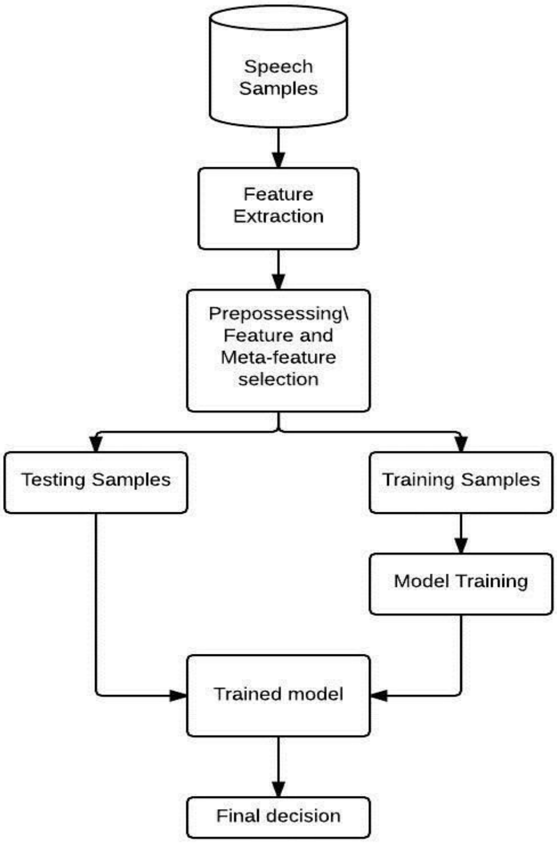

Figure 1-2: Pattern Recognition Steps

dimensional data. However, judgments that are based on performance and complexity cannot be generalized everywhere, because the classification models' characteristics are not completely independent of the previous steps, including feature extraction and selection. For example, LDC draws cluster borders based on a multivariate Gaussian distribution of the data belonging to that cluster, and performs well when applied to de-correlated data. Feature selection techniques, which are able to provide a de-correlated subset of attributes, are more appropriate for classification methods like the LDC, like the meta-feature selected by PCA. Therefore it is shown in Chapter Six how PCA with LDC scheme can outperform other schemes like the PCA with SVM. The performance of any pattern recognition task can be improved by the fusion of more than one model, whether in terms of the score level or the classification level. The decision of the models can be weighted to highlight the influence of some models over others as in Figure 1-3. In emotion recognition, a speaker-independent classification is more applicable. And mainly the researchers adopt the Leave One Speaker Out (LOSO) cross validation approach.

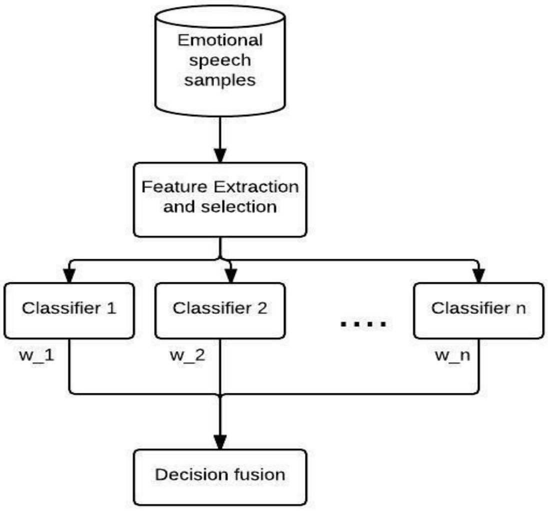

Figure 1-3: Fusion at classification level

In this thesis we investigate and compare the performance of various classifiers and various fusion schemes in terms of accuracy rates. However, we shall also investigate the pattern of performance of such classifiers for different feature selection and feature reduction schemes.

1.4 Contributions of this thesis

This thesis aims to investigate major issues regarding emotion recognition in speech, especially developing and testing the performance of a number of SER schemes. We shall use innovative approaches to feature extraction, feature selection and reduction, and classifiers.

We started by investigating new sites within the speech signal to extract features relevant to emotion. In Chapter Four, a set of features, which are extracted from the LP-residual signal, is suggested for emotion recognition. These features include:

1. Mel Frequency Cypstrum Coefficient (MFCC) of the LP-residual signal.
2. Linear Predictive Cypstrum Coefficient (LPCC) of the LP-residual signal, and
3. Wavelet Octave Coefficients of Residual (WOCOR).

These sets of features are used for speaker recognition, but are not used in emotion recognition. The suggested features show the ability to improve the accuracy of recognition when fused with a large number of emotional features (6,652) extracted by OpenEar software (Eyben, et al., 2009).

In this thesis we also investigate the controversy about a small size feature set relative to emotion, and consequently suggest the use of a large number of features in SER studies. Meta-feature extraction and selection is proposed in this thesis as a solution to avoid the curse of dimensionality, including the use of Data Independent PCA (DIPCA), where PC projection is computed by data samples independent from the samples involved in training the classification, and the random projections that are based on compliance with the new innovative paradigm of Compressive Sensing (CS). These techniques also avoid model complexity without affecting performance. We suggest a number of CS-compliant Random Projection (RP) matrices for transforming data into a meta-feature space instead of PCA. The advantage of RP matrices over PCA is their non-adaptive nature (i.e. the independence of the data used for model training), which results in significantly lower computational costs.

The investigation presented in Chapter Five shows how the RP can be used for emotional space adaptation from speaker influence, together with an SVM classifier. However the SVM have some serious drawbacks, including the need for balancing the number of samples for each pair of classes, and building a set of machines for multi-class models, as well as high computation cost. An alternative classifier is investigated in Chapter Six using the fact that the de-correlated meta-features generated from a high dimensional feature space using PCA is suitable for classification, based on normal multivariate distribution. Therefore LDC using pooled estimated covariance for data pre-processed by PCA outperforms the well-known SVM classifier and the state of the art results.

We have also proposed an ensemble classification model that extracts meta-feature subsets using PCA with “well-selected” weights and have resulted in improved recognition accuracy. For further improving the performance of the SER scheme, we designed a multi-level classification by combining some of the confused classes. The confused classes have been detected using confusion matrices and a Non metric Multi-Dimensional Scaling (NDMS) that visualize the similarities between classes. A comparison with what we are aware of state of the art schemes shows that most of the models suggested in this thesis outperform state of the art models in terms of model performance.

A score-based classification has been used to show that the speech sentences somehow encapsulate a mixture of emotions, such that the same speech sample can reveal different emotions. Classifying speech samples confidently is rarely found in non-acted datasets, and stronger in the professionally acted datasets, especially in the case of datasets that are designed for analysis or synthesis and not for recognition purposes such as the EMO-Berlin dataset. This thesis investigates the presence of more than one emotion in an individual speech sample.

1.5 List of publications

Currently a part of this thesis is presented in two publications:

1. An article submitted to Computer Science and Electronic Engineering Conference (CEEC) 2013, entitled "*Excitation source and low level descriptor features fusion for emotion recognition using SVM and ANN*". The details of the contribution of the publication are presented in Chapter Four.
2. An article submitted to the SPIE 9497, Mobile Multimedia/Image Processing, Security, and Applications 2015 conference, entitled "Emotion Recognition in Speech: Tools and Challenges".

1.6 Thesis Organization

The thesis includes seven chapters starting with the introduction in Chapter One followed by a background on emotion from a psychological point of view, and the common features extracted from speech signals in Chapter Two. A literature review on emotional datasets and speech features used for SER is also available in two different sections in Chapter Two. In Chapter Three, feature pre-processing techniques are presented, together with the classification methods used. Research regarding feature pre-processing and classification are also reviewed in Chapter Three.

The first contribution of this thesis regarding features extracted from the LP-residual samples is presented in Chapter Four. This is followed by the investigation of meta-feature extraction and selection using different forms of PCA and RP using SVM classifier in Chapter Five. In Chapter Six the use of an LDC classifier is investigated to overcome some of drawbacks of SVM. The data is also pre-processed by some forms of the PCA and RPs. Meta-features subsets extracted by PCA is also fused in this chapter to improve recognition performance, in which the assumption of the availability of a mixture of emotion in the same portion of speech is studied. Finally the conclusion of the whole thesis is come in the last chapter.

Chapter Two
Emotion and speech signal features

Emotion is known to modulate a variety of human-machine modes of interaction. Audio-visual modes including speech and facial expression are among the most commonly researched modes of interaction for detection and identification of emotions. Other detectable emotion-related signals include hand and body gesture as well as internal physiological changes. In this thesis the focus is on using speech alone for the detection and recognition of emotion.

SER (Speech Emotion Recognition) is based on two major aspects, the first one is the meaning of emotion and the second one is the kind of speech signal features or parameters that could be relevant to emotion in speech. As a pattern recognition task, pre-processing the emotional speech data by extracting and then selecting the suitable features is followed by classification. A background of the feature extraction step is covered in this chapter.

In this chapter a proper description of emotion from a psychology point of view will be presented in the first section (1.1), to understand how it could benefit in designing an automatic emotion recognition system. In order to highlight the difficulties involved in SER, we review the literature on the databases designed for emotion recognition from speech signal in section (1.2). In section (2.3) the most well-known and important types of speech signal features, which are reported to be relevant to emotion is shown. The final section (2.4) will cover a literature review of emotional features extracted from speech signal.

2.1 How psychologists view human emotion?

Understanding the exact meaning/manifestation of emotion is necessary in any emotion related applications. Psychologists conducting research into human emotions need emotional data/samples to be in a format that includes labelling of the samples by accurate emotions. Such needs raise questions like how many emotional categories should be considered, how to collect emotional data, and how to map the variety of emotional categories to manageable lists of labels.

Psychological studies are concerned with factors such as identifying the significant and fundamental list of emotions, how and why are they differentiated, and the situations of emotion eliciting. This would have a clear impact on any emotion related application. Understanding the emotion helps in recognizing it, predicting its influence on other behaviour, and controlling it. However, emotion definition is a matter of opinion amongst the psychologists, regarding the root of the emotion, basic and derivative emotions, and other aspects. The "biology versus culture" debates about the root of emotion attracted many researchers. On one side the traditional theory by Darwin claims the emotion to be survival-related pattern that have evolved to help the adaptation with the species environment in the evolution stages; while many anthropologists and social psychologists attacked the Darwin's theory; as they attributed the emotion to sociocultural factors (Steidl, 2009). In this study we adopt the concept suggested by Schuller et al (Schuller, et al., 2011b), who followed the definition of Cowie, (Cowie, et al., 2001b) that '*emotion*' is what is "*present in most of life but absent when people are emotionless*"; this is the concept of *pervasive emotion.* This concept includes both the prototypical emotions like (anger, happiness…etc.), and what is called emotion-related states, in addition to the all related concepts to emotion like "*emotional intelligence".*

2.1.1 Emotion categories and dimensions

In pattern recognition applications, labelling the data to classes is essential. The model will be trained using some samples labelled to what is assumed to be the correct classes. It is obvious that labelling the emotional data depends on the theory behind the emotion and the observer skill. Defining and predicting the number of emotions face serious difficulties amongst the psychologists. However, categorization of emotions can be done in both dimensional and discrete emotion models.

2.1.1.1 Dimensional categorization

Emotions can be distinguished according to some characteristics like: arousal, valence, and dominance. Emotion is possible to be distinguished using one-dimensional or multidimensional model.

2.1.1.2 Uni-dimensional Model

Some researcher argues about the usefulness and sufficiency of one dimension categorization of emotion. The dimension can be the *arousal/excitation* (low to high), or the *valence* (unpleasant to pleasant). The valence dimension is reported as the most important representation, which represents the principal of emotion. It is now one of the most accepted criteria for emotion and effect studies (Scherer, 2000). Distinguishing emotions to Positive and Negative is an example of one-dimensional categorization.

2.1.1.3 Multidimensional Model

Emotion are also represented in the literature by a multidimensional model. The suggested dimensions are *Pleasantness-unpleasantness* (valence), *rest-activation* (arousal), and *relaxation- attention* (dominance) (Cowie, et al., 2001b). Multi-dimensional approach provides a theoretical model that could be compatible with the Opponents Processing in Emotion sensing (Solomon , 1980).

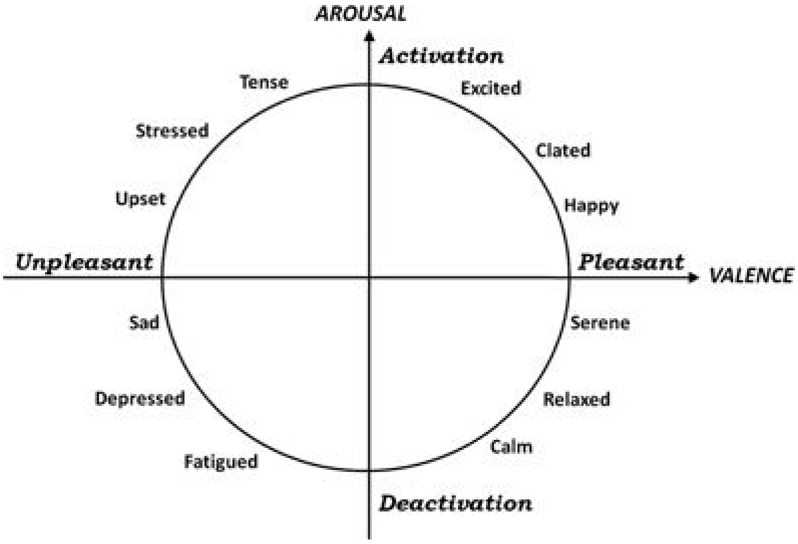

Figure 2-1: Arousal and valence dimensions in emotion representing.

Multidimensional representation helps in improving the capability for improved separation between emotions that are close in terms of arousal, valence or dominance. The obvious benefits from using two dimensions relate to the ability of visualizing the differences between emotions in an illustration (see Figure 2.1). In two/three dimensional space the Euclidian distance could also be used to measure the differences (Steidl, 2009). However there is no consensus about the exact needed number of emotions.

Table 2-1: A Selection of Lists of "Basic" Emotions (Ortony & Turner, 1990)

Reference	Fundamental emotion	Basis for inclusion
Arnold (1960)	Anger, desire, despair, fear, hate, hope, tendencies love, sadness	aversion, courage, dejection, Relation to action
Ekman, Friesen, & Ellsworth (1982)	Anger, disgust, fear, joy, sadness, surprise	Universal facial expressions
Frijda (personal communication, September 8, 1986)	Desire, happiness, interest, surprise, wonder, sorrow	Forms of action readiness
Gray (1982)	Rage and terror, anxiety, joy	Hardwired
Izard (1971)	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Hardwired
James (1884)	Fear, grief, love, rage	Bodily involvement
McDougall (1926	Anger, disgust, elation, fear, subjection, tender-emotion, wonder	Relation to instincts
Mowrer (1960)	Pain, pleasure	Unlearned emotional states
Oatley & Johnson- laird (1987)	Anger, disgust, anxiety, happiness, sadness	Do not require propositional content
Panksepp (1982)	Expectancy, fear, rage, panic	Hardwired
Plutchik (1980)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological processes
Tomkins (1984)	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	Density of neural firing
Watson (1930)	Fear, love, rage	Hardwired
Weiner & Graham (1984)	Happiness, sadness	Attribution independent

2.1.2 Discrete emotion model

Another Model of categorization emotion is to label each individual emotion alone as a single state. Many researcher tries to define basic emotions, which are supposed to be expressed by mammals as a result of many theories mostly extending Darwin's theory of emotion evolution. Although there is a no agreement about the exact number of basic emotions, the most popular list is the Ekman's categorization of basic emotions. He regarded Anger, disgust, fear, joy, sadness, and surprise as basic emotions. Table 2-1 shows some other suggestions of the basic emotions. This diverse categorization is a result of the disagreement of understanding the emotion concept itself, and would certainly add to the challenge of automatic emotion recognition. Defining basic emotion is important to explain some routine observation about emotion, which appears in human and some higher animals as well (Ortony & Turner, 1990). However recent studies tend to bypass the use of basic emotions and instead consider the alternative concept of Emotion-related states (Batliner , et al., 2011). In this thesis the investigation will include both basic emotion and emotion-related states.

2.2 Emotional Speech Data Collection

Emotion could be present in (or detected from) different human modes of interaction with the environment and is manifested by expressions in the face, changes in speech tones, gestures, and/or electroencephalogram (EEG). Consequently emotion related datasets are collected by different researchers depending on their choice of interaction mode. For example, face expressions datasets include faces in different expressions (Kanade, et al., 2000), whereas SER datasets are collected from speech, which contain emotional speech samples (Engberg & Hansen, 2007; El Ayadi , et al., 2011). Multi modal datasets is also suggested including face images and speech signals in the form of videos as in (Douglas-Cowie, et al., 2005).

Being focused on speech emotion recognition, the main datasets investigated in this thesis are collected in benchmark databases that contain speech samples for sufficiently large sample of the population. Emotional speech applications use these datasets to train suggested models and/or to analyse the emotional characteristic included in the speech signal. Designing or eliciting an emotional data set, is a serious challenging task for the SER research community, due to difficulty in assessing how a recorded dataset is natural and/or usable. The prompted data is frequently criticized for being difficult to create a natural setup for expressing different emotions (El Ayadi , et al., 2011; Batliner , et al., 2011; Steidl, 2009; Schuller, et al., 2011b). Characteristic differences of detected ranges in contour fundamental frequency and other speech parameters values between real life clear emotion sentences and sentences acted by professional actors is reported by (Williams & Stevens, 1972). The actors are asked to act the emotions included in the pre-designed structure of the data set, which might lead to produce a “full blown” emotions.

Another issue regarding the prompted data is the prototypical approach in fixing the emotion included in the database, which lead to the absent of many available emotions or emotion-related states in the real life. In addition to that the prompted emotional data is not applicable when linguistic features are used, because the adopted texts are somehow predefined and fixed independently of any emotion. Deciding whether the acted recordings are really well representing the emotion or not is another challenge that faces the prompted data. Human examiner (annotators or experts) are used to judge and score emotional samples. Some database designers remove the samples that are not representing the emotion well (Burkhardt, et al., 2005). This kind of sentence selection might also influence the emotional data and raise doubts about the recording being natural. However some of these databases (like DES (Engberg & Hansen, 2007), Emo-Berlin (Burkhardt, et al., 2005), SUSAS (Hansen & Bou-Ghazale , 1997)) were not collected for SER, but for quality measuring of emotional content syntheses. In contrast the prompted data is easier in design, and unlike the non-prompted datasets, number of samples per class is controlled/pre-defined.

The non-prompted data is recorded in an unsupervised environment, which means that the subjects are not directed to express a specific emotion, but they are led to an emotional state, and then the produced emotion is recorded without their knowledge. This kind of data seems to be more ‘realistic’. However, some researchers agreed to call this kind of data as a non-prompted emotional data, instead of name it as spontaneous or natural. (Schuller, et al., 2011b).

Recording non-prompted samples of emotional data also faces many serious challenges. In the non-prompted speech data, some emotion-related states might appear, in which not suggested initially, but rather is likely to be available in the real life. The emotion-related states (like emphasized, rest, positive, etc.) are not necessary to be listed under the prototyped emotion categorization or the basic emotions.

As we discussed above, the number of samples per emotion/class is not pre-defined in the non-prompted data, therefore it depends on emotion related states available in the recordings. Additionally the non-prompted data needs another effort to label the speech samples to each individual emotions or emotion-related states. Experts are usually involved to manage this process. Sometimes some of the emotion-related states appear rarely in the collected data, which encourage mapping the available emotion-related states to less number of classes. For instance (Steidl, 2009) suggests a heuristic approach to map many emotional related-states to less number of categories. State of the art studies encourage the use of the non-prompted data, because it is more realistic and helpful in designing a real life application (Batliner , et al., 2011).

The emotional databases are also different in terms of the target that the dataset is designed for. Some emotional databases are designed infant-directed like the dataset designed by (Slaney & McRoberts, 1998), which collect 500 samples from adults talking to their infants and classified 65% of the samples correctly. More other database is adult-directed as shown in the next section.

2.2.1 Historical review of emotional databases

The history of speech emotional databases started in the late 1990s with few databases but mostly limited in the number of samples (about 500), number of enrolled subjects (about 10), and number of uttered sentences (about 10). For example, the DES database (Engberg & Hansen, 2007), contains 419 speech utterances, expressed by 4 professional actors (2 male and 2 female) who were acting 5 emotions, using 9 different sentences. The emotions that were meant to be expressed by the speakers in this database were *anger, happiness, neutral, sadness, and surprise*. The recordings are judged by twenty native speakers, whereby 67% are correctly evaluated. The Emo-Berlin database (Burkhardt, et al., 2005), contains studio recording of 10 sentences uttered repeatedly by 10 subjects (5 male and 5 female) who were asked to express 7 emotions (*anger, happiness, neutral, sadness, fear, disgust, and bored*). There were 20-30 judges participating afterwards in evaluating and labelling the emotion recordings. About 500 out of 800 utterances are evaluated by 20 subjects as appropriate emotional sentences with minimum score of 60%, assignable with minimum of 80%, and correct labelling accuracy rate of 84.3%. The Emo-Berlin database is one of the databases that will be used in this thesis.

Expansion of the size of the databases in terms of samples, subjects, and modals, in addition to adding more natural life characteristics to the database is followed in designing the later databases. For example, the eNTERFACE database (Martin, et al., 2006) uses 42 subjects (9 female) and 1227 samples. To add more cultural diversity factors the subjects were chosen from 14 nations (see Table 2-2). Another naturalize attempt in this database was leading the subjects into an emotional situation by preparing them to listen to an emotional story and then they asked to utter five appropriate sentences per each emotion. The emotions induced in this database are *anger, disgust, fear, joy, sadness, and surprise*.

The first attempt to create a "spontaneous" data set was made through the SUSAS database (Hansen & Bou-Ghazale , 1997), where a predefined set of words captured from English air-commands. A total of 3593 samples under different level of stress are produced. This was followed by the design of a comprehensive spontaneous database, called FAU Aibo database (Batliner, et al., 2008a;Steidl, 2009).

The FAU-Aibo database (Steidl, 2009;Batliner, et al., 2008a), was designed by recording children's sound, which are coloured by different emotion, when they interact with Sony's pet robot Aibo. The children were led to believe that the robot is responding to their commands, whereas it was actually controlled by a human operator in a Wizard-Of-Oz manner. Five experts labelled each word in the database independently, into 10 categories: angry, touchy, joyful, surprised, bored, helpless, motherese, reprimanding, emphatic, and 'other' for the remaining cases. The categories were mapped into four classes*: anger, emphatic, neutral,* and *positive* in addition to the fifth class for *rest*. The Aibo corpus formed the focus of the Interspeech 2009 emotion challenge (Schuller, et al., 2011a). Aibo dataset is one of the datasets that adopted in this thesis, and more details about it will be given in chapter 4.

For further details a comprehensive review on databases can be found in (Schuller, et al., 2009b; El Ayadi , et al., 2011; Koolagudi & Rao, 2012b). Figure 2-2, shows the percentage of number of emotions used in 32 datasets reviewed by (Koolagudi & Rao, 2012b)) while Figure 2-3 shows the number of simulated, elicited, and non-prompted datasets. The number of non-prompted datasets is just 25%, and the 2-class datasets is the more present in the available emotional datasets.

Table 2-2: Geographic distribution of participants in the eNTERFACE database (Martin, et al., 2006)

Country	Number of Subjects	Country	Number of Subjects
Belgium	9	Cuba	1
Turkey	7	Slovakia	1
France	7	Brazil	1
Spain	6	USA	1
Greece	4	Croatia	1
Italy	1	Canada	1
Austria	1	Russia	1

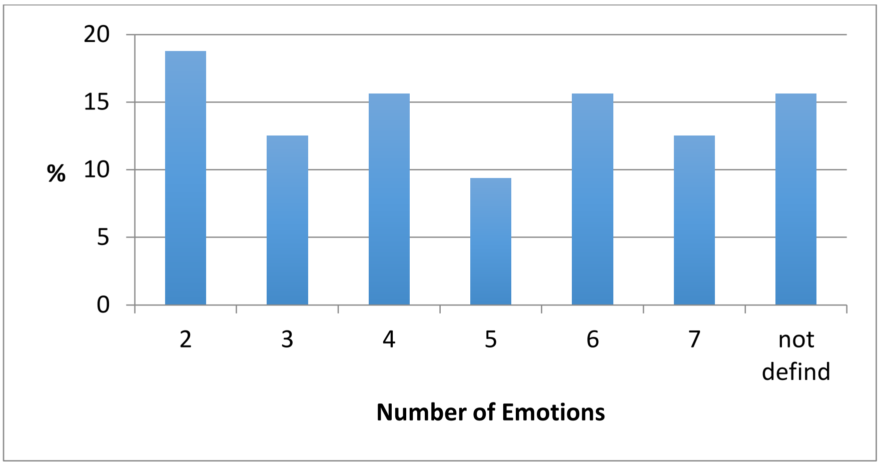

Figure 2-2: The percentage of emotions from 32 datasets collected in (Koolagudi & Rao, 2012b)

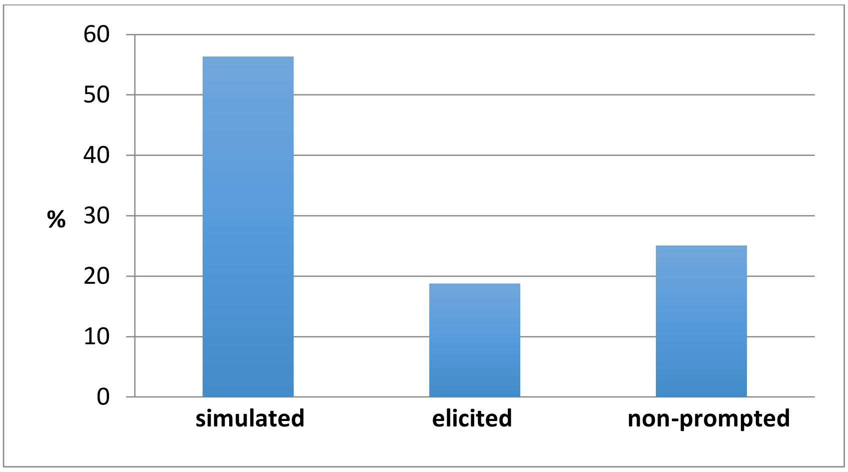

Figure 2-3: The percentage of dataset types from 32 datasets collected in (Koolagudi & Rao, 2012b)

2.2.2 Shortcoming of emotional databases

State of the art studies encourage using non-prompted (spontaneous) data sets due to their similarity to the real life human emotional states. Still the non-prompted data sets have two major problems, which could be taken into account when generating emotional data. The first issue is regarding the neutrality of the monolog form of emotional speech that adopted in the available data sets. This shortcoming happens because of the absence of emotion-related state that accompany the dialog conversation. The difficulty of preparing the recording situation for spontaneous emotional dialog conversation is a serious challenge, but necessary to bring the emotional datasets furthermore towards the real life environment.

The second issue is that the universality assumption (albeit implicit) of collected datasets as representing a wide range of cultures presents yet another serious research challenge due to the fact that no existing emotional dataset can claim such another characteristic of existing. In the next section a brief review on some cultural influence on emotion is presented.

2.2.3 Cultural influence on emotion

Some researcher as a consequence of the argument about the linguistic, has studied the influence of culture on emotion expressed by people brought up in different cultures, and culture influence verses the universal characteristic of emotions. Generally, studies report the universality of emotion to some extent. For instance in (Elfenbein & Ambady, 1986) listeners' ability of recognizing emotions from different cultures, has been claimed to lead recognition accuracy score much higher than the random probability chance of recognition. This kind of claims are used to some extent as an indicator of the universality of emotion in different cultures. In contrast some other studies reported the influence of linguistic prior knowledge that judgers have on the emotional contents. The comprehensive study by Scherer et al. (Scherer, et al., 2001) show that recognition accuracy score of native judgers of German language emotional contents range from 74%- 84%, while Indonesian judgers of the same recording achieved only 52% accuracy. These two seemingly different investigations seems to indicate that the emotional characteristic is universal to some extent, but the linguistic characteristic has also its role in understanding the emotional contents in a text. We believe that this argument needs more investigations to settle this kind of questions.

2.3 Acoustic Feature extraction for SER

Feature extraction is one of the most important steps in pattern recognition (if not the most). In quantitative pattern recognition studies, extracted features are usually used to build a mathematical model governing the variability of these features. Extracting the appropriate SER features reflects the available knowledge about emotion characteristics as well as the influence of the person's emotional state on the speech signal. In this study we deal with acoustic features of the speech signal that extracted using different approaches.

In general, features are either selected depending on a pre-knowledge based approach or "brute force" approach". In pre-knowledge based feature extraction, experience plays a role in fixing the number of targeted features to feed the classification model. Practically, the number of features in a pre-knowledge SER system tends to be limited to not more than few hundreds. While in the "brute force" approach, feature extraction aims to cover as much as possible characteristic that the speech signal has, hopping to capture as much as possible information about emotion embedded in the speech signal, although this may result in the presence of redundancies.

The vocal tract system produces speech from a time varying signal with a time varying excitation. Therefore the speech signal is non-stationary in nature, while most signals processing tools assume time invariant system and excitation, i.e. stationary signal. Therefore these tools are not directly useful for speech processing, and mainly short time parameter estimation is more applicable. Therefore, to overcome this issue of non-stationary over a long utterance of speech, it is customary to divide speech signal into frames through which the signal is almost stationary (Rabiner & Schafer, 2007). In the rest of this section a background of feature sites will be covered.

2.3.1 Prosodic features

Prosodic refers to some speech signal characteristics like stress, rhythm, and intonation, which are reported to reflect the emotional states, as well as the intonation of the speaker which change the meaning of sentences forms like question, or command. In this section major prosodic parameters, will be presented.

2.3.1.1 Pitch and fundamental frequency

Although the pitch and fundamental frequency are different in the sense that pitch refers to perceptual characteristic, while fundamental frequency is physical parameter of the speech signal, but there are an agreement about the correlation of both of pitch and fundamental frequency. Fundamental frequency (F_0) changes through a spoken sentence, offer information about the intonation and stress happens along spoken words and sentences. F_0 is measured in Hertz according to the fundamental period of vocal closure. F_0= 1/T, where T is the fundamental period.

Due to the non-stationary characteristic of the speech signal F_0 is usually computed in the short-term signal. To have more clear view of the changes happen of the F_0 along the speech signal sample, F_0 envelop parameters (like slope, onset, offset, etc.) is computed, beside other statistics like minimum, maximum, mean, median, etc. of F_0 value along the whole signal (Buckow, et al., 1999). Figure (2-4) shows some statistics computed from the pitch contour. In this work we follow the implementation of openEAR toolkit (Eyben, et al., 2009), which use the cepstrum and Autocorrelation based algorithms for F_0 computation.

2.3.1.2 Energy and zero crossing based features

Like the F_0, short-term energy is computed for each frame, and the statistics is applied for the whole utterance. Unlike the fundamental frequency some statistics like the minimum, onset, and offset of the energy envelope, does not make sense because the minimum energy is usually zero or close to zero (see Figure 2-5).

The short time energy is computed as follow (Rabiner & Schafer, 2007):

$$E_n = \sum_{i=-\infty}^{\infty} (x(i)w(n-i))^2 \qquad (2.1)$$

Where *w* is hamming window.

While the zero crossing is computed by:

$$Z_n = \sum_{i=-\infty}^{\infty} 0.5|sign(x(i) - sign(x(i) - 1)|\, w(n-i) \qquad (2.2)$$

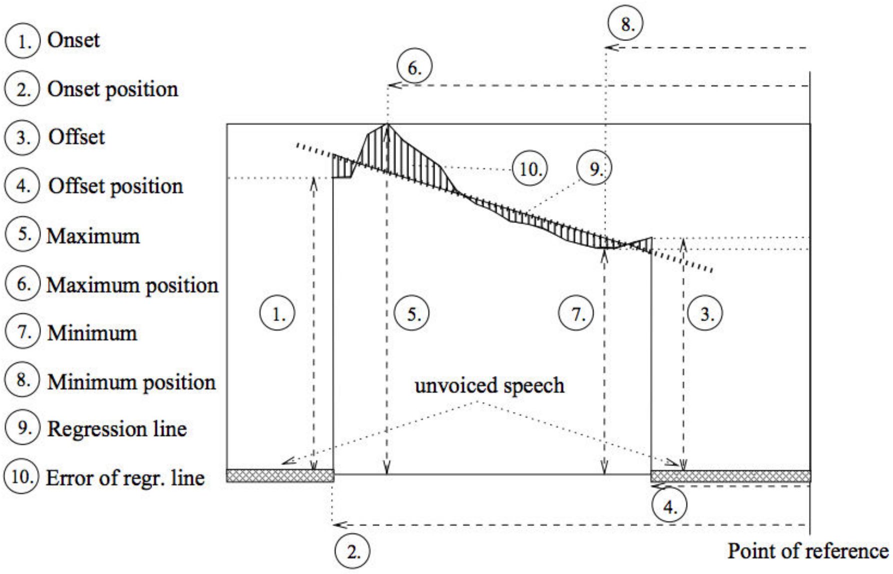

Figure 2-4: Example of features used to describe a pitch contour (Buckow, et al., 1999)

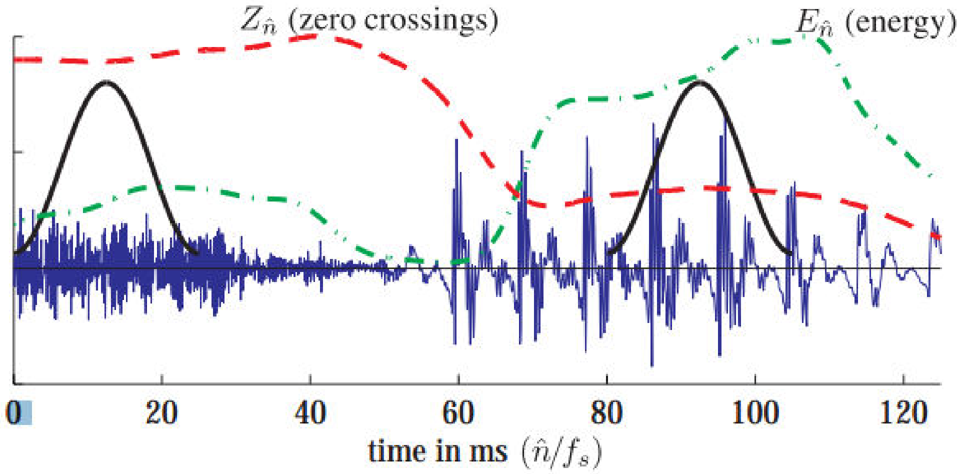

Figure 2-5: Section of speech waveform with short-time energy and zero-crossing rate superimposed

2.3.1.3 Duration and Pauses features

The duration of the words, and the normalized duration by the number of syllables is the core of this set of duration and pauses features. This kind of feature is usually computed by manually segmentation of the words. For full automatic emotional data analysis the computation of this kind of feature is not necessary. Speaking rate is also reported as duration related features (Muto , et al., 2005), consequently, the position of maximum and minimum of F_0 and the energy, beside the onset and offset of F_0, are used for this purpose.

The duration of pauses, whether it is filled (like “Emm”, “Ahhh”, etc.) or silent pauses, is considered as pauses-based features (Steidl, 2009).

2.3.2 Spectral Features

Spectral information reflects the distribution of waveform frequencies along the speech signal. Spectral features are reported to carry information on text contents, speaker’s identity, and emotional state (Paliwal, 1998;Pérez, et al., 2012). Formant frequencies are widely used spectral features that characterize the phones in a speech signal, especially the lower formants. Speaker-related information can also be found in the higher formants. Furthermore frequency bands magnitudes, spectrum Roll-off and centroids are all computed to represent the spectral features. We first describe some of these spectral features.

2.3.2.1 Formants

Formants of a speech signal represent the resonance that happen while the generated airwaves pass through the vocal tract. To model the speech resonance characteristic, the vocal tract is modelled by a simple tube of length L (about 170mm), this tube is normally divided into N equal length sub-tubes S_i with different widths, (see Figure 2-6).

To understand how the formant is estimated, it is necessary to refer to Fant’s suggested model of speech production (Fant, 1970) as shown in (Figure 2-7). The passage of the speech signal, generated by the vocal cords, through the vocal tract is subject to multiple filtrations. This system is modelled by the following Z-space filter equation:

$$F(z) = U(z).V(z).R(z) \qquad (2.3)$$

Here, U represent the glottal pulses scaled by the voiced controller, V is the vocal tract filter, and R represents the lip radiation filter. This system can model the speech production as a linear time invariant system.

The disruption of signal flow happened with the transition from one cylinder to another. Thus V(z) filter is defined by the following formula:

$$V(z) = \frac{K}{1 - \sum_{i=1}^{N} a_i z^{-1}} \qquad (2.4)$$

where K is an acoustic flow parameter of the signal. And a_i is the ith linear prediction coefficient.

The function V(z) has N/2 pairs of complex conjugate poles (i.e. the zeros of the polynomial in the denominator) as shown by the following expression:

$$1-\sum_{i=1}^{N} a_i z^{-1} = \prod_{i=1}^{\frac{N}{2}} 1-2e^{-c_i T}\cos(b_i\,T)z^{-1}+e^{-2c_i T}z^{-2} \qquad (2.5)$$

The spectral envelop peaks known as the formants. The LP- spectrum has an amplitude frequency peak at frequency $b_i/2\pi$, which represent the formant frequencies. The formant bandwidth is also estimated as $c_i/2\pi$ (Rabiner & Juang , 1993).

There are many features related to the formants like, formant frequencies, bandwidth, their amplitude, onset and offset of envelop between a pair of two formants.

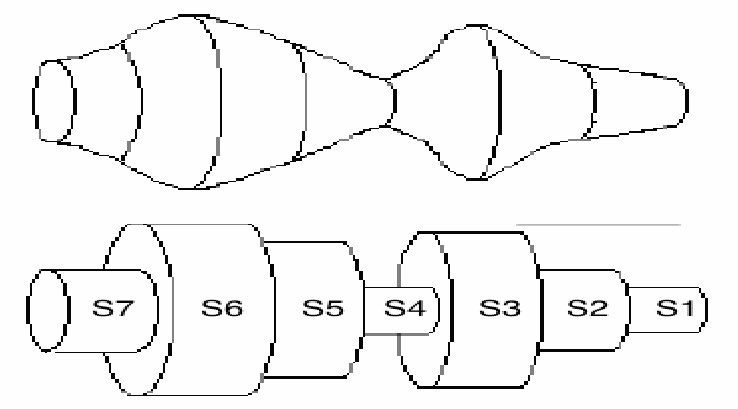

Figure 2-6: simulation of the vocal tract tube.

Statistical parameters of the distribution of the magnitudes of different frequency bands over the length of the speech signal are used as equivalent features to formant features. In this work spectral magnitude peaks in different bands, their positions, bandwidth, and slope parameters are also used beside many different energy bands magnitudes.

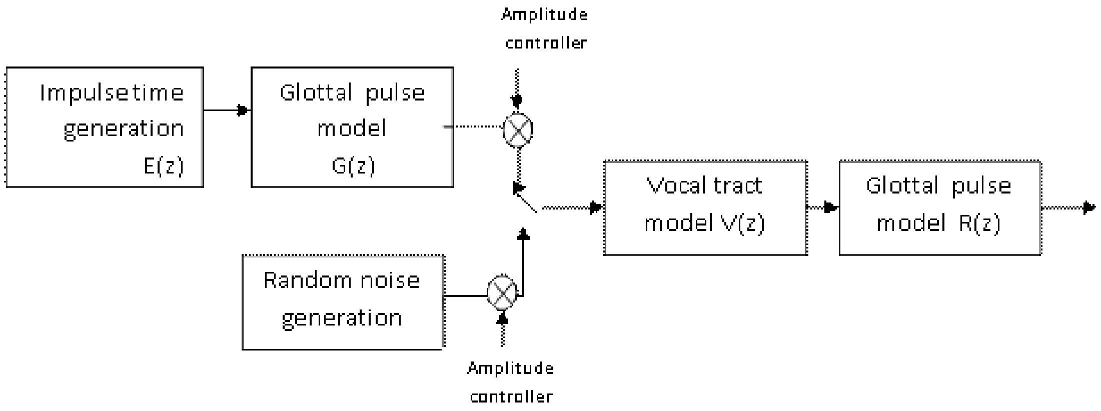

Figure 2-7: Speech Production model

2.3.2.2 Fourier transform based feature

Fourier transform (FT) is the most common way of analysing signals into their spectral sub-bands. However, the non-stationary characteristics of the speech signal recorded over a relatively long time make the use of FT of limited benefits. This is due to the fact that FT cannot provide simultaneous information on the involved frequencies and their location within the signal. But it is well known that speech signal tends to be stationary in short frames of about 16 ms and hence the use of short-term Fourier analysis provides the appropriate analysis tool (Rabiner & Juang , 1993). To conserve the continuity of the signals across the frames especially in the beginning and the end of the frames overlapped windows (like hamming window) that extenuate the amplitude in the frame sides. Many features can be extracted from the Fourier spectrum of the feature like the spectrum centroid, and Flux. The spectrum centroid is computed from the Fourier transform of the signal as follows:

$$F(k) = \left\| \sum_{n=1}^{N} x(n) \times e^{-i2\pi k\frac{n}{N}} \right\| \quad (k = 1, 2, \dots, N) \tag{2.6}$$

where $x(n)$ is the input speech signal of length N, and $F(k)$ is the amplitude of the spectrum. While, spectrum centroid is defined by the following equation:

$$S_c = \frac{\left(\sum_{k=1}^{\frac{N}{2}} k \times F(k) \right)}{\left(\sum_{j=1}^{\frac{N}{2}} F(j) \right)} \tag{2.7}$$

The spectrum Flux F_S^t is computed as follows:

$$F_S^t = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\frac{F^t(k)}{E^t} - \frac{F^{t-1}(k)}{E^{t-1}}\right)^2} \tag{2.8}$$

where E^t is the energy of the frame *t*. (Chen , et al., 2012)

2.3.2.3 Mel Frequency Cepstrum coefficients (MFCC)

The MFCC is a representation of the short-term power spectrum of the voice signal. The computation of the coefficient is based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of the frequency.

The MFCC's can be computed using the algorithm shown below:

1- Pre-emphasize the signal by apply a high pass filter. This is required to obtain similar amplitude for all formants, such a filter has the following representation in Z domain:

$$H(z)=1- a\ z^{-1}.$$

2- Compute the Fourier transform of the signal and find its log power spectrum.

3- Convert the frequencies into mel-scale, according to the following formula:

$$m = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \tag{2.9}$$

4- Create a band pass filter bank and find the summation of Db power spectrum of passed frequencies for different bands.

5- Apply the cosine transform on the values obtained to reach the spectrum coefficient of the spectrum of the nonlinear male scale frequency.

2.3.3 Voice quality features

The quality of the voice (e.g. breathiness or harshness) heard by others is an influencing factor in determining the speaker's emotion, and therefore parameters that measure voice quality must be included in feature investigation for SER. Voice quality features are partly based on the pitch and intensity parameters, and reflect the characteristics of the glottal resonance. Computing of the fundamental frequency and the intensity for cycle-to-cycle variation is one way of estimating some of the voice quality features like jitter and shimmer.

The relative jitter measurement of the voice quality is estimated as follows:

$$jitter_r(i) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T(i+1) - T(i)|}{\frac{1}{N}\sum_{i=1}^{N}|T(i)|} \tag{2.10}$$

Where $T(i)$ is the wavelength of the fundamental frequency F_0, also known as the F_0 period, and N is the number of extracted periods of F_0.

While the shimmer value satisfy the following equation:

$$shimmer_r(i) = \frac{\frac{1}{N-1}\sum_{i=1}^{N}|A(i+1) - A(i)|}{\frac{1}{N}\sum_{i=1}^{N-1}|A(i)|} \tag{2.11}$$

where, $A(i)$ are the extracted peak-to-peak amplitude data and N is the number of extracted fundamental frequency periods (Farrús, et al., 2007).

2.3.4 Linear Predicted analysis

The mathematical representation of Fant's speech production model, which was explained in 2.3.2.1 has the following transform function:

$$\frac{\mathrm{S(z)}}{U(z)} = \frac{\mathrm{G}}{1 + \sum_{k=1}^{q} \mathrm{a'_k\ z^{-k}}} \tag{2.12}$$

Where q is the number of parameters used in generating the speech signal.

The filter above could be written in the time domain as:

$$\mathrm{s[n]} = \sum_{k=1}^{p} \mathrm{a_k}\ s[n-k] - Gu[n] \tag{2.13}$$

In reality it is very difficult to find the parameters of this filter directly, but the idea is to design a filter to estimate the current speech sample from the previous samples by using the stricter of the speech production model, represented by:

$$\frac{\mathrm{E(z)}}{S(z)} = 1 - \sum_{k=1}^{q} \mathrm{a'_k\ z^{-k}} \tag{2.14}$$

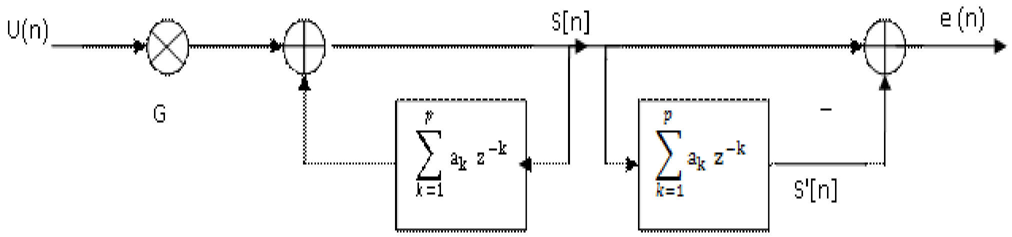

Figure 2-8: Filter of Generating estimated speech signal from Linear Predictive coefficients

The estimated samples are used to define the mean absolute error with the actual sample as shown in (Figure 2-8).

From equation 2.14 and 2.12, we obtain the following equation:

$$\frac{E(z)}{U(z)} = G \, \frac{1 - \sum_{k=1}^{q} \alpha_k \, z^{-k}}{1 - \sum_{k=1}^{q} a_k \, z^{-k}} \tag{2.15}$$

If $a_k = \alpha_k$ then:

$$\frac{E(z)}{U(z)} = G \quad \text{i.e.} \quad e(n) = Gu(n)$$

It is expected that the prediction error *e(n)* would be large for voiced speech at the beginning of each pitch period. The summation error is defined as follows:

$$E = \sum_{n=1}^{N} e(n)^2 = \sum_{n=1}^{N} \left(s(n) - s'(n)\right)^2 \tag{2.16}$$

$$E = (\sum_{n=1}^{N} s(n) - \sum_{n=1}^{N} a_k s(n-k))^2 \tag{2.17}$$

The parameters a_k are predicted as the values that minimize the expected error value E. This can be determined by differentiating the equation with respect to a_i.

$$\partial E/\,\partial a_i = (\sum_{n=1}^{N} 2\,[s(n) - \sum_{n=1}^{N} a_k s(n-k)] \;\; (-s(n-i)) \tag{2.18}$$

where N is the total number of the samples. To minimizing E, set $\partial E/\,\partial a_i = 0$, i.e

to minimize the parameter E, 2.18 is equalized to zero

$$\sum_{n=1}^{N} s(n)\, s(n-i) - \sum_{n=1}^{N} a_k \sum_{k=1}^{p} s(n-k) \quad s(n-i)=0 \tag{2.19}$$

P is the number of the parameters to be estimated. Now,

$$\sum_{k=1}^{p} a_k \sum_{n=1}^{N} s(n-k)s(n-i) = \sum_{n=1}^{N} s(n)s(n-i) \qquad (2.20)$$

This could be simplified in terms of the Autocorrelation function R as:

$$\sum_{k=1}^{P} a_k \quad R(k-i) = R(i) \qquad i = 1,2,\ldots,P \qquad (2.21)$$

The corresponding system of equations:

$$\begin{bmatrix} R(0) & R(1) & R(p) \\ R(1) & R(0) & R(-p+1) \\ \cdots & \cdots & \cdots \\ R(p) & R(-p+1) & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(0) \\ R(1) \\ \cdots \\ R(p) \end{bmatrix}$$

Solving this linear system equation above result in estimating the parameters $a_{i,}$ which is called linear predictive coefficients (Rabiner & Juang , 1993).

The LP-residual signal $r(n)$ of the speech signal is the difference between the original signal $s(n)$ and the predicted signal $\hat{s}(n)$ obtained by Linear Prediction:

$$r(n) = s(n) - \hat{s}(n) \qquad (2.22)$$

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \qquad (2.23)$$

2.4 Emotion relevant speech features

Over the last two decades there has been a vast amount of researches in the area of emotion recognition from speech. The above list of speech based features have been incorporated into different types as emotion discriminating feature vector representations for SER schemes and their performances were tested on different emotional speech datasets with varying degrees of success. This section will cover a literature survey of some works regarding the relevant features to emotion extracted from the speech signal.

At the early stages of emotion recognition researches, the focus was on using prosodic features (Pitch, duration and energy/ intensity) but much less attention was given to other features like voice quality features (Schuller, et al., 2011b). The prosodic features is still investigated and included in the state of the art SER studies. For instance (Batliner , et al., 2011) investigated 6 different sets of features relevant to emotion all in which include pitch related features, 5 of them include energy based features, and 4 include the duration features. Later on the amount of features have been extended to include prosodic, spectral, and voice quality features, in such a way that “brute force” approach of feature extraction (thousands of feature) were proposed frequently (Schuller, et al., 2009b;Hassan & Damper, 2012).

The discovery of more new sites and features is an ongoing process. Recently some new sets of features extracted from the LP- residual signal have been added to the list of investigated features. For instance, (Chauhan, et al., 2010) use 40 samples of voiced speech LP-residual signal and another 40 samples around the glottal closure instants. While (Koolagudi & Rao, 2012b) uses features of the epoch parameter extracted from the glottal closure region of the LP-residual signal. These studies were motivated by the relevance of these features to emotion, as well as the limited use of the LP-residual signal in emotion recognition. In this study we will propose a yet another set of features to be extracted from the LP-residual, which has not been used before in emotion recognition (see Chapter four).

In the state of the art, for feature extraction, there are trends toward extracting acoustic features using many sites with many parameters (Batliner , et al., 2011). However such a high dimensional data need to be followed by dimension feature selection, or reduction step. This high number of features is a consequence of many factors but primarily the non-agreement on a limited number of emotion-related features (Vogt & Andre, 2005). For instance, studies on the relationship between global prosodic speech features and the basic emotions have shown that prosodic features provide a reliable indication of emotions. Moreover, many contradictory remarks/conclusions are reported in the literature on the effect of emotions on prosodic features. For instance, while Murray and Arnott (Murray & Arnott, 1993) indicate that a high speaking rate is associated with the emotion of anger, Oster and Risberg (Oster & Risberg, 1986) makes opposite conclusions. In another study Yang et al. (Yang & Lugger, 2010) argued that the prosodic features could separate classes in the arousal dimensions, whereas voice quality features are effective for discriminating classes in the valence dimension. In a similar study by Ebyen et al.

(Eyben, et al., 2010a), mel-spectrum is reported to be the most effective feature to separate classes in the valence dimension. However, in a correlation based feature selection study, Perez et al. (Pérez, et al., 2012)argued that the most relevant features to discriminate classes in the valence, arousal, and dominance dimensions are MFCC, cochleagrams, and LPC respectively. They also point out that Mel features have significant discriminatory contributions in valence and dominance, while energy features have discriminating characteristics in the arousal dimension. A strong relation between voice quality and perceived emotion with listening human subjects is demonstrated in (Gobl & Chasaide, , 2003). Voice quality is also reported regularly in reference to the full-blown emotion (i.e. acted) (Cowie, et al., 2001a).

The non-agreement on a set of features could be due to the nature of emotional databases used (Batliner , et al., 2011), which are captured under different conditions and factors. Simulated versus spontaneous, number of classes, cultures, or the circumstance of recordings could all be contributing to the need of different sets of features.

Emotional features are extracted using different speech portions. Features could be extracted for each word and then statistics is applied for chunks or sentences. This kind of approach is called two layer based feature extraction. In contrast the single layer is extracting features from the whole chunk (Steidl, 2009). In order to extract features from the supra-segmental level of the speech signal, some researchers model the dynamic changes occurring in the speech signal using feature values in short segments, and hidden Markov Models have been used for this purpose (Daniel , et al., 2006). Another approach is to project the sequence of segmental features (for each limited duration) through different statistics to generate a set of global features for each speech sample. Recently, researchers adopted a ‘brute force’ static set of features by extracting a set of Low Level Descriptors (LLDs), and computing many functionals of them, which tried to solve the non-agreement of the more important features relevant to emotion. This work adopts the use of 6552 acoustic features extracted by OpenEar software (Eyben, et al., 2009).

The OpenEAR software extracted features based on the OpenSmile project (Eyben, et al., 2010b), in which extracting 250k features is implemented in real time factor rtf 0.044. This performance encourages the use of high number of features in SER. The study of Koolagudi (Koolagudi & Rao, 2012b) containes a comprehensive review on different features used in emotion reconition.

And finally, Schuller et al. (Schuller, et al., 2007) showed that pooling together features extracted at different sites did improve emotion recognition accuracy performance. However, this pooling increase the dimension significantly leading to the curse of dimension, which is traditionally dealt with by dimension reduction schemes that create smaller sets of meta-features. While using "brute force" approach of feature extraction has been highlighted in some study (Schuller, et al., 2009b;Hassan & Damper, 2013), we are not aware of any research along these lines. This would be one of the main tools that we will be investigating for SER.

2.5 Summary

Chapter 2 covered a background material on emotion in a psychological point of view, to provide the context in a guiding manner for conducting research in SER. This is also useful in designing more emotion relevant data sets so that more “natural” emotional content could be captured in the collected data. Some efforts for creating emotional datasets has also been surveyed in this chapter, highlighting the challenges that one face even in collecting reliably annotated data. The main quantitative emotion–related features and parameters extracted from speech signal, have also been described and presented in this chapter, including most of the OpenEAR based extracted feature that is adopted in most all of the experiments (a list of the OpenEAR features will be presented in Chapter 5). The chapter ended with a brief literature review of SER research work that use various extracted sets of features.

The next chapter will cover another major part of the theory background regarding the SER covering feature pre-processing and classification, before reporting the work done in the rest of the thesis.

Chapter Three
Feature Pre-processing and Classification

The traditional process of pattern recognition starts with extracting features/attributes that relate to some variable aspects of the objects/patterns under investigation, followed by selecting the "most relevant" features or meta-features that are used in training the classification model. In chapter 2, we presented various types of features, which are reported to be relevant to emotion. The non-agreement on "limited" set of emotional feature is a serious challenge that leads to the adoption of a "brute force" large number of features. To improve the classification model performance and/or reduce the complexity, feature or meta-feature selection is a widely used and reasonable solution. We shall present in section (3.1) some pre-processing methods; in order to generate a feature subspace that would include a "proper" lower dimension representation of the data, to be input into the SER model. Evaluating the performance of the SER model will have to be based on an appropriately chosen classification technique. In section (3.2) a description of the classification techniques that are investigated in this thesis, will be described. Finally a literature survey on both topics is included in sections (3.3) and (3.4).

3.1 Pre-processing methods

Speech signals generate many attributes/features that can be used to reveal information about different human characteristics and the semantics of the speech (eg. speaker, spoken text & context, and emotion). Knowledge about various sets of features relevant to emotion, helps in extracting a compact representation of speech that is relevant to emotion recognition and the performance and\or efficiency of the adopted SER model. Data pre-processing stage aims to identify and collect the most informative features/meta-features as an input into an appropriate classification model. It is essential to recognize that in practical applications, informative set of features/meta-features does not necessarily preclude the presence of hidden redundancies or correlations.

This section presents some well-known and common dimension reduction techniques that are meant to preserve similarities of the data samples in terms of the relevant objects/concepts. Particularly, we shall describe in the next two subsections (3.1.1 and 3.1.2) methods for feature and meta-feature selection.

3.1.1 Feature selection

Feature selection is a common pattern recognition stage and is much more effective in a non-high dimensional feature space. It is the process of selecting a subset of features based on a criterion\objective. The objective is either measuring some characteristics of the data attributes by filtering (Filters) OR choosing a proper subset of features that yield the highest accuracy rate post classification. The latter approach is referred to in the literature as a Wrapper (Kojadinovic & Wottka, 2000). Filter methods are simple, efficient, and use data characteristic like the correlation or mutual information to select the "suitable" feature subset. Wrapper methods use a predetermined classifier to train, evaluate and determine the selected features (See Figure 3-1).The cost of optimization when selecting a suitable subset among a high number of features, is relatively high, especially when using the wrapper methods. In contrast the filter methods is known to be less accurate than wrappers (Yu & Liu, 2003). This is due to the fact that in the training stage Filter methods are independent of the choice of classifier, while wrappers use the same classifier to optimize the objective cost function and evaluate the classification model (Bermejo, et al., 2012).

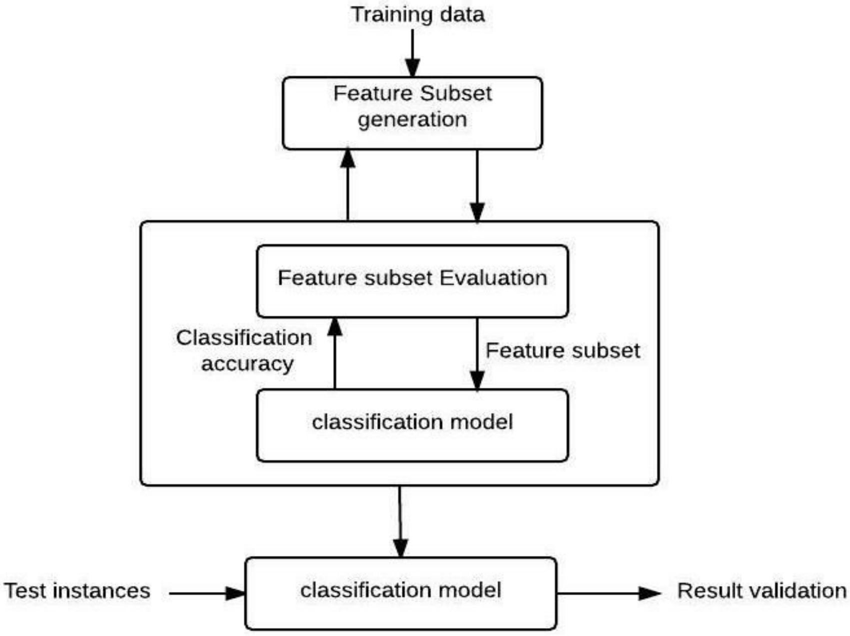

Figure 3-1: A diagram of wrapper feature selection procedure.

Ideally, feature selection methods should de-correlate the feature space besides seeking the relevant features. A popular example on feature selection methods is the Sequential Floating Feature Selection (SFFS) (Pudil, et al., 1994), which has the ability not just to add only or remove only features (like Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS)), but also to drop some other features with high criterion error, by including a back step of removing non-suitable features during feature addition (Schuller, et al., 2011b).

SFFS is initialized with an empty set of feature, and then adds features based on a specific criterion (objective function), which is equivalent to one step in SFS. Subsequently, at each stage the worse feature in terms of the same performance criterion, is removed equivalently to one step of SBF. Therefore SFFS dynamically update the set of feature until reaching the “optimized” feature set. Unlike the PCA and Linear Discriminant Analysis (LDA) projections, SFFS attempts to optimize its objective function rather than de-correlating the features although the back step may reduce correlation. In other words, SFFS does not guarantee selecting the optimum de-correlated feature subset (Batliner, et al., 2008b).

For high dimensional data, there are some different proposals for feature selection. These methods are based on the data-sparsity and aim to solve the problem in terms of l_0 minimization, which is equivalent to the cardinality of the set of non-zero coefficients. However, due to the fact that l_0 minimization is a No-deterministic Polynomial-time (NP) hard problem, the l_1 minimization is adopted to have an approximated solution for the same problem. For instance, in (Dehua, et al., 2013) an iterative way of feature selection is embedded inside the SVM classifier by exploiting the l_1 minimization for both the SVM and feature selection optimization. In chapter 5, we shall use an iterative version of sparse-SVM and present the obtained results for SER.

3.1.2 Meta-feature selection

Meta-features are simply a linear combination of atomic features and therefore are obtained by linear transformation of the original data vector space, i.e. determined by multiplying the data vectors by projection matrices. Ideal meta-features are expected to preserve the samples distances/similarities, i.e. orthogonal matrices. These projections might be obtained by supervised training using the training data samples, like the LDA, or unsupervised but depend on the training data like the Principal component Analysis (PCA), it can also be unsupervised and independent of any data like Random Projection (RP). A number of meta-features are selected from the transformed space to obtain a manageable and informative representation of the data.

3.1.2.1 Principal Component Analysis (PCA)

In general data variables are mostly correlated to each other, unless the structure of the data is not simple or the dimension is small (Jolliffe, 2002). For any application with sufficiently large number of d-dimensional sample data, the PCA is a change of base linear transformation of R^d for which there is small set PC of base vectors, called principal components, such that the projection of almost all samples onto base vectors not in PC are small and negligible (See Figure 3-2). Consequently, PCA can be used a as a dimension reduction tool, by representing the samples in the subspace of PC. The PC base vectors are the eigenvectors of the covariance of the data (after subtraction of their mean vector). The eigenvalue λ corresponding to the eigenvector v of the data covariance C is a measurement of the variance in the direction of v, and usually only small number of these eigenvalues have significant absolute values (Mika, et al., 1999). Eigenvectors corresponding to the relatively small number of significant eigenvalues determines the columns of the PCA projection that de-correlates the data samples.

The PCA algorithm:

1. Let $\boldsymbol{X} = \{x_1, x_2, \dots, x_N\},\ x_i \in R^d$, and N is the number of samples of input data, and suppose that $\boldsymbol{m}$ is the d-dimensional mean vector.
2. Define $\boldsymbol{Y} = \{y_1, y_2, \dots, y_N\}$, where, $y_i = x_i - \boldsymbol{m}$**.**
3. Compute the symmetric covariance matrix $\sum$ of **Y**,
4. Compute the eigenvectors of $\sum$, as columns of the PCA transformation matrix T of order k.
5. Transform the data X into $X^{'}$ using the following formula:

$$X' = X\,T \tag{3.1}$$

X' is the set of transformed data, referred to as the Eigen data, in the principal component sub-space.

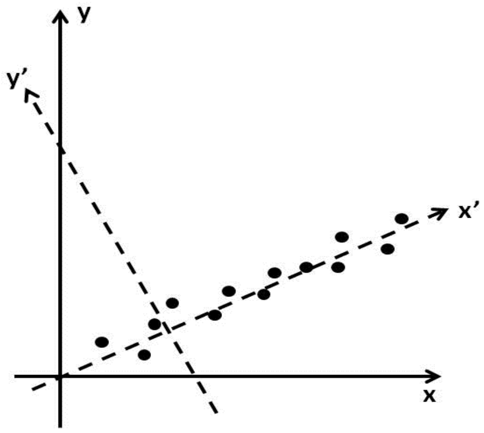

Figure 3-2: The Principal component x' represent the direction where data has high variance.

Applying PCA is used to produce a de-noised version of the data by discarding a proper number of eigenvectors with lower eigenvalues. PCA serves as general-purpose tools with various applications like information extraction, dimension reduction and data visualization (Moon & Phillip, 2001;Yeung & Ruzzo, 2001).

3.1.2.2 Linear Discriminant Analysis (LDA)

Unfortunately, the construction of the PCA projection does not take into account class information, and trained with sample data belonging to different classes, consequently within class samples could be dispersed while across class samples could get closer. The LDA is designed to avoid this situation by solving a generalized eigenvalue problem (i.e. of the form (A– λ**B**) **v**= 0) whose eigenvectors, corresponding to the most significant eigenvalues, form the columns of a projection matrix that maintaining the class discriminations as much as possible (see Figure 3-3). For that purpose Fisher (Fisher, 1936) suggested the objective presented in equation (3.2), which aims to maximize the between classes and minimize the within class distances, represented in the following objective function:

$$O(W) = \frac{W^T S_B W}{W^T S_W W} \tag{3.2}$$

here, S_B and S_W are the between classes and the within classes scatter matrices respectively, and W is the vector variable. If N in the number of classes, then:

$$S_B = \sum_{i=1}^{N} (\mu_i - \mu)(\mu_i - \mu)^T \tag{3.3}$$

$$S_w = \sum_{i=1}^{N} \sum\nolimits_{x \in C_i} (x - \mu_i)(x - \mu_i)^T \tag{3.3}$$

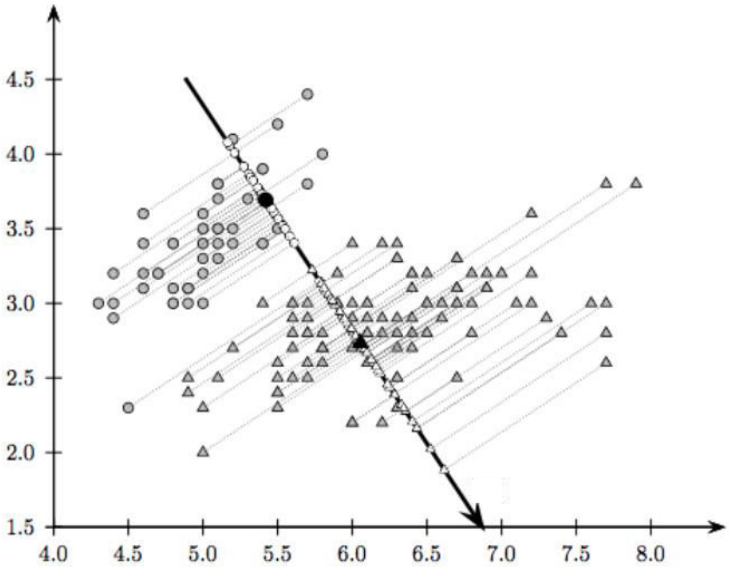

Figure 3-3: Linear Discriminant Directions

Where, μ_i is the mean of the samples in the i-th class, and μ is the mean vector of the whole data samples.

3.1.2.3 Random Projection (RP)

RP projection matrices $P_{d\times k}$ that have orthonormal columns offer a non-adaptive alternative that project n observations each in a d-dimensional space $X_{n\times d}$ into k-dimensional subspace with k<<d using random entries selected by a specific distribution, such that most of the sample distances/similarities are preserved, or subject to minute errors. The transformed data $X_{n\times k}^{RP}$ are define by the formula:

$$X_{n\times k}^{RP} = X_{n\times d} \times P_{d\times k} \tag{3.5}$$

The computation of random projection matrices is not as expensive as the PC projection, for instance the order of complexity for projecting the data $X_{n\times d}$ on a random matrix $P_{d\times k}$ is *O(nkd).* The idea behind using random matrices is the Johnson-Lindenstrauss lemma (Johnson & Lindenstrauss, 1984), states that under certain conditions a high dimensional data can be transformed into a random subspace; thereby the distances are preserved, with an error allowance ε.

Johnson-Lindenstrauss lemma

Given $0 < \varepsilon < 1$, a set X of N points in R^d, and a number $k > 8\,ln(N)\,/\varepsilon^2$, there is a linear map f: $\mathrm{R}^\mathrm{d} \rightarrow \mathrm{R}^\mathrm{k}$ such that:

$$(1-\varepsilon)\|x-y\|^2 < \|f(x)-f(y)\|^2 < (1+\varepsilon)\|x-y\|^2 \tag{3.6}$$

Designing a suitable random matrix P is an essential issue. For example a high dimensional Gaussian random matrix is not necessarily orthogonal in its columns, which means that projecting into this sub-space might lead to data distortion. However, Hecht-Nielsen (Hecht-Nielsen, 1994) has shown that in high dimensional space there exist many nearly orthogonal directions beside the original directions. In (Bingham & Mannila, 2001) the orthogonality of matrices was measured by the difference $||P^TP - I||$between the P^TP and the identity matrix I. For the Gaussian $d \times k$ matrix it has been shown the difference is at most *1/k*. Moreover practical results show that any binary element content matrices are also suitable a random projection matrices (Shuhong, et al., 2010). In other words, these authors remove the need for checking orthogonality for such random matrices.

Random projections are closely related to Compressive Sensing (CS) in the sense that latter aims to reconstruct original signals through sensing a compressed version of the signal, using Restricted Isometric Random Projections (RIRP).

3.1.2.4 Compressive sensing (CS)

The concept of CS is a recently developed concept that relaxes the stipulated stringent requirement of Nyquist-Shannon sampling theory on the number of samples needed for the perfect recovery of a signal. According to the CS paradigm, sparse (or nearly sparse) signals can be recovered well from far less than double the highest frequency, as required in the signal, required by the Nyquist-Shannon Theory. In most signal/image applications, signals are compressed and represented by sparse feature vectors using transformations (e.g. wavelet & DCT) that remove significant correlations and redundancies in the capture signals. Compressive sensing is an attempt to recover the sparse (compressed) signal directly by sensing a relatively small number of linear measurements (i.e. meta-features) of the signal/image from which the significant (i.e. the nonzero) sample/pixel values of the compressed signal/image can be recovered. CS relies on the sparsity and the incoherence principals that must be satisfied by the projection matrix. The sparsity means that the information content of the signal can be represented by less than the original signal bandwidth, i.e. the signal depends on a number of degree of freedom much less than the signal length.

While the incoherence characteristic means that the compressed version of the signal can be sensed or acquired back (Emmanuel & Wakin, 2008). The main crucial observation in compressive sensing is that the design of sensing and sampling methods/dictionaries with the ability to capture adequate information of the signal, is possible with relatively smaller number of linear meta-feature measurements than the dimension of the data vectors. Additionally, these dictionaries (sensing matrices) meanwhile capturing information are not trying to comprehend the signal, i.e. non-adaptive. Such a dictionary should satisfy what is called Restricted Isometric Property (RIP) (Candès & Wakin, 2008), closely related to the above Lemma.

3.1.2.5 Restricted Isometric Property (RIP)

Let Ω denote the set of all length-N vectors with K non-zero coefficients. An MxN measurement matrix Φ has the restricted isometric property (RIP) with parameters K and $\delta \in (0,1)$, if it satisfies:

$$(1-\delta)\,||x||^2 \leq ||\,\Phi x||^2 \leq (1+\delta)||x||^2 \quad \text{for all } x \text{ in } \Omega \tag{3.7}$$

The definition above indicates that any Gram matrix of any RIP matrix has eigenvalues in [1-δ, 1+δ]. The RIP is sufficient but not necessary to guarantee the recovery of the sparse signals, i.e. there might be other non-RIP matrices that can reconstruct sparse signals. The random matrix, which has the Gaussian distribution N(1,0), or a Bernoulli distribution satisfy the RIP condition. The RIP random matrices are called Random Restricted Isometric Projection (RRIP) matrices.

3.1.2.6 Examples of RP matrices

The Bernoulli and Gaussian random matrices are known to be RRIP, but these are by no means the only cases. Generally any independently and identically distributed (i.i.d.) data matrix could be used for this purpose. Below we present some other suggested RRIP matrices:

1-Achlioptas (Achlioptas, 2003), suggests a matrix defined as:

$$A_c(i,j) = \begin{cases} +1 \ \ with\, probability\ of\ 1/6 \\ 0 \ \ \ with\, probability\ of\ 2/3 \\ -1 \ \ with\, probability\ of\ 1/6 \end{cases}$$

2- Circulant Toeplitz matrix, where every row in this matrix is the right cyclic shift of the row above, as below:

$$T_n = \begin{bmatrix} a_0 & a_1 & \cdots & a_n \\ a_n & a_0 & \cdots & a_{n-1} \\ \dots & \dots & \dots & \dots \\ a_1 & \dots & a_n & a_0 \end{bmatrix}$$

3- Matrices in the form, below can also be used, (e.g. k=2, and $\epsilon = 0.1$):

$$C = \begin{bmatrix} \frac{1}{k} & \frac{1}{k^{1+\varepsilon}} & \frac{1}{k^{1+2\varepsilon}} & \cdots & \frac{1}{k^{1+(n-1)\varepsilon}} \\ \frac{1}{k^{1+(n-1)\varepsilon}} & \frac{1}{k} & \frac{1}{k^{1+\varepsilon}} & \cdots & \frac{1}{k^{1+(n-2)\varepsilon}} \\ \dots & \dots & & \dots & \dots \ \dots \\ \frac{1}{k^{1+\varepsilon}} & \frac{1}{k^{1+2\varepsilon}} & \frac{1}{k^{1+3\varepsilon}} & \cdots. & \frac{1}{k} \end{bmatrix}$$

The columns in these matrices are not required to be orthonormal as in the case of the PCA projection matrix, but it must be linearly independent. This property might pose the random projection as more suitable technique in some cases, especially when the features are not highly correlated.

A special form of RP matrices, which is frequently used here, is the Binary Random matrix (BR) whose entries are {0, 1} with equal probability of *r* =0.5. The column elements of the matrix will be the coefficients of the linear combinations forming the projected meta-features. The matrix entries here, will either ignore some of the attributes scores or will not change its value, which might be useful to pick a suitable random representation of the data in lower dimensions without scaling the individual sample dimension score. This version of random binary matrix is not RIP. It shares the duality of elements with the RIP version of the Bernoulli matrix that has elements of the form $\pm 1\sqrt{n}$, with a ratio of r=0.5 for each. Many other random binary matrices are used as projections matrices (Shuhong, et al., 2010).

3.2 Classification and validation

Supervised classification is based on training the classifier using a labelled set of samples, and evaluate the model's performance with another independent set. The aim of training the classifier is to generate separators between different classes' clusters. The separator might be a hyper-plane (linear subspace), which is optimized based on the position or the distribution of the training samples, or it could be a combination of hyper-planes (non-linear). One of the most popular classifier is the k-NN (Cover & Hart, 1967), which classifies a sample based on the nearest neighbour sample(s). Usually the measurement of nearness is the Euclidian distance, and majority voting of the label of the k nearest samples will be responsible in making the final decision. In this section we introduce some classifiers, which are used in this thesis, like the Support Vector Machine (SVM) (Cortes & Vapnik, 1995), Linear Discriminant Analysis (LDC), and Artificial Neural Network (ANN).

Training a given set of data influences the separator hypothesis produced by the classifier, which might lead to the problem of over-fitting or a bias decision toward the data samples participating in the training stage. To overcome this problem, the adopted evaluation process influences the classifier performance. The most well-known evaluation approach is the cross-validation techniques, which aims to let as much as possible data participating in training and evaluating the model, so that the achieved accuracy rate is more or less achievable for independent data. In this section a description of some classifiers and validation techniques will be displayed.

3.2.1.1 The Support Vector Machine (SVM)

The SVM classifier aims to find an optimal hyper-plane, which separate two classes' samples. Optimality is obtained by maximizing the width of the margin between the hyper-plane and data points (support vectors) on the border of both classes' clusters.

Given N samples data $X = \{x_1, x_2, \dots, x_N\} \in R^d$, and a set $Y = \{y_1, y_2, \dots, y_N\}$, where $y_i \in \{-1, 1\}$, is the class label of x_i.

The targeted separator an affine hyper-plane f in R^d is defined as follow:

$$\{x: f(x) = x^T\beta + \beta_0 = 0\} \tag{3.8}$$

Here, β is a unit vector normal to the hyperplane and β_0 is a constant vector.

The induced classification rule by $f(x)$ is simply defined by:

$$G(x) = sign[x^T\beta + \beta_0] \tag{3.9}$$

Perfect class separation is equivalent to having $y_i f(x) > 0, \forall\, i$. However, the fact the training set may not be fully representative of the real-life application then avoiding over-fitting and bias, SVM classification decision will be made on the bases of a sufficiently large margin M (see Figure 3.4 a), needed to separate data points from the hyperplane, i.e. solve the following optimization problem:

$$\max_{\beta,\beta_0} M \tag{3.10}$$

$$subject\ to{:}\ y_i\ (x^T\beta + \beta_0) \geq M, \qquad \mathrm{i} = 1,2,\ldots,\mathrm{N} \tag{3.11}$$

Replacing M with $\frac{1}{||\beta||}$, yields the equivalent minimization problem:

$$\min_{\beta,\beta_0} ||\beta|| \tag{3.12}$$

$$subject\ to\ y_i\ (x^T\beta + \beta_0) \geq 1, \qquad \mathrm{i} = 1,2,\ldots,\mathrm{N} \tag{3.13}$$

which is a convex, quadratic criterion with linear inequality constraint.

SVM introduces slack variables $\xi = (\xi_1, \xi_2, \ldots, \xi_N)$ to solve this problem to allow modelling inseparable data samples (see Figure 3.4 b). The sum of the ξ_i's need to be bounded by a constant K, because > 1 leads to misclassification of the point x_i. Thus the final optimization has the form:

$$\min ||\beta||\ subject\ to \begin{Bmatrix} y_i\ (x^T\beta + \beta_0) \geq 1 - \xi_i\ \forall i \\ \xi_i > 0, \qquad \sum \xi_i < constant \end{Bmatrix} \tag{3.15}$$

For computation purposes the equation (3.15) is rewritten in the following form:

$$\min_{\beta,\beta_0} \frac{1}{2} ||\beta||^2 + C \sum\nolimits_{i=1}^{N} \xi_i \tag{3.16}$$

$$subject\ to\ \xi_i > 0,\ y_i\ (x^T\beta + \beta_0) \geq 1 - \xi_i \quad \forall i \tag{3.17}$$

This optimization is solved using the Lagrange function:

$$L_P = \frac{1}{2} ||\beta||^2 + C \sum\nolimits_{i=1}^{N} \xi_i - \alpha_i [\, y_i\ (x^T\beta + \beta_0) - (1 - \xi_i)] - \sum\nolimits_{i=1}^{N} \mu_i \xi_i \tag{3.18}$$

L_P is minimized by equating its gradients w.r.t $\beta,\ \beta_0, and\ \xi_i$ to 0. For more details on how to solve this optimization problem see (Hastie, et al., 2001, pp. 132-135)

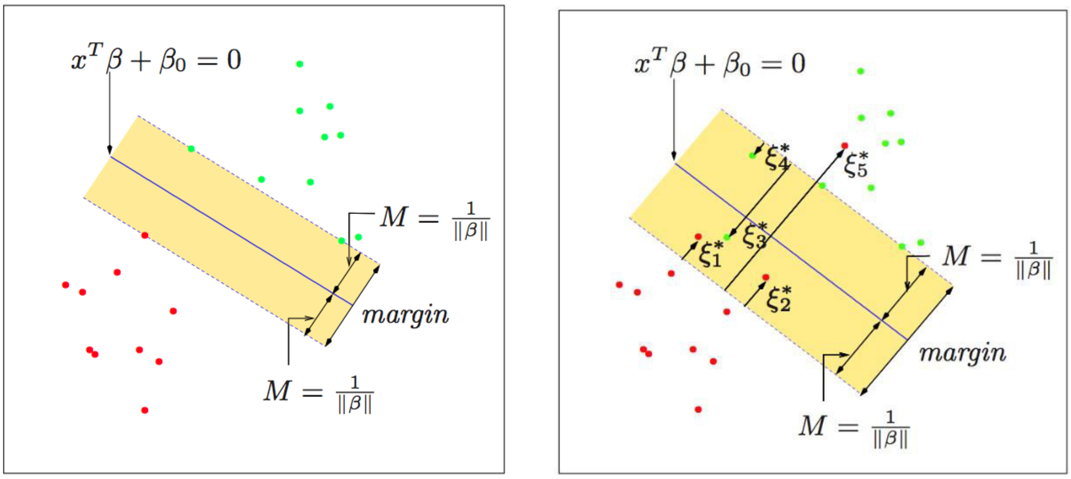

Figure 3-4: (a) Separated classes (b) overlapped classes

3.2.2 Linear Discriminant Classifier (LDC)

The LDC classifier assumes that the population $\pi_i\ of\ the\ class\ C_i, where\ i = 1, 2, \dots, K$ of each class follows a multivariate Gaussian distribution:

$$f(x|\pi_i) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\ e^{\left[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right]} \tag{3.26}$$

Where μ_i is the mean vector of the class C_i and samples are classified to class π_i with largest value of $p_i f(x|\pi_i)$ where the monotonicity of the log function implies that log $(p_i f(x|\pi_i))$ is equivalent to $p_i f(x|\pi_i)$. Here, p_i is the prior probability of the class i. The linear score function is re-expressed as follows:

$$S_i = -\frac{1}{2}\ \mu'\Sigma^{-1}\mu + \mu'\Sigma^{-1}x + \log p_i \tag{3.27}$$

$$S_i = a_{i0} + \sum_{j=1}^{K} a_{ij}x_i + \log p_i \tag{3.28}$$

where:

$$a_{i0} = -\frac{1}{2}\ \mu'\Sigma^{-1}\mu \tag{3.29}$$

$$a_{ij} = jth\ element\ of\ \mu'\Sigma^{-1} \tag{3.30}$$

The priori probability of population i is estimated based on prior knowledge of the class' distribution. In this study we substitute the population covariance by the pooled covariance C_p of the whole classes' data. Thus the score function representing the classification rule is set as follows:

$$S_i(x) = -\frac{1}{2}\mu' C_p^{-1}\mu + \mu' C_p^{-1} x \tag{3.31}$$

3.2.3 Artificial Neural Network (ANN)

ANN aim to find a set of weights corresponding to a diagram of nodes that map the input data into their targets through a hidden layer(s) and an output layer. The set of weights will be adjusted in an iterative process to minimize the Mean Square Error (MSE) defined by:

$$E = \frac{1}{2}(t - y)^2 \tag{3.32}$$

where t is the actual target value of the sample, and y is the output value of the model.

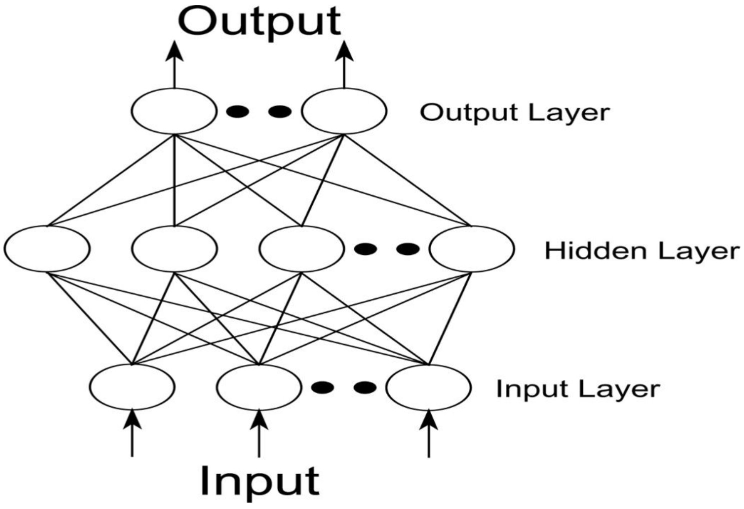

Figure 3-5: An example of Neural Network topology.

In chapter four we adopt the use of feed forward back-propagation neural network. ANN adopts various kinds of topologies depending on the application (e.g. see Figure 3-5). The error defined in (3.32) can be minimized by different optimization techniques, like the Gradient decent methods.

The input sample $X = (x_1, x_2, \dots, x_N)$ feeds into the N input layer nodes in first step, the value of at each node in the hidden layer will be calculated through the output of function fed by linear combination of the input layer nodes value with the initial weights corresponding to the hidden layer:

$$net_i = f(WX^T + b) \tag{3.33}$$

The same procedure will be repeated to produce the value at the nodes in the next layers until the output nodes has their values. The output values y and the target values t is used to calculate the error E defined in (3.32). Weight adjusting based on gradient decent method for the activation function:

$$y = \frac{1}{1 + e^{-z}} \tag{3.34}$$

can be defined as:

$$\frac{\partial E}{\partial w_i} = \Delta\, w_i = (y - t)y(1 - y)x_i \tag{3.35}$$

where, $\Delta\, w_i$ is the amount of change when updating w_i in each iteration. To decrease the chance of over-fitting, a learning parameter α is multiplied by the $\Delta\, w_i$. Thus $\alpha\ \Delta\, w_i$ characterize the weight-adjusting rule in back-propagation ANNs. The adjusted weights are the parameters of the trained classification model, which are used in evaluating the model using the test samples.

3.2.4 Ensembles

Ensembles are machine learning algorithms that uses more than one classification models; the overall decision is obtained from a weighted\un-weighted majority voting of the included classification decisions (Dietterich, 2000). Ensemble may outperform individual classifiers, if the decision of the classifiers in the ensemble disagreed on some individual samples and the classifier errors are not correlated (Dietterich, 1997). The classifier models can have different decisions on the same sample, if they are trained by different versions of data (Bagging) or classified by different classifiers.

Researchers are adopting different ways for producing various versions of data from the same source. One such approach is to randomly choose more than one subset of instances (with replacement) from the original set of features. In other words suppose that we have the data matrix $X_{n\times d}$ of n samples each in R^d, then the k subsets of data can be defined as:

$$X^i_{m\times d}, \text{where m} < \text{n}, \qquad \text{and i} = 1{,}2, \dots, \text{k}.$$

Alternatively choose more than one subsets from the feature set, to train independent models to have their own decision. Hence, k subsets of data is defined as follows:

$$X^i_{n\times p}, \text{where p} < \text{d}, \qquad \text{and i} = 1{,}2, \dots, \text{k}.$$

In this thesis we adopt also choosing a number of meta-feature subsets produced by PCA to be classified by different models collected in an ensemble scheme.

Final decisions are made by majority voting, i.e. if the classes' labels denoted as c_i, and the ensemble consists of K models then the number of classifiers that predict the instance x as the class c_i is given as:

$$v_i = \left|\left\{M_j(x) = c_i \middle| j = 1{,}2, \dots, K\right\}\right|, where\ K\ is\ the\ number\ of\ classifiers$$

Predicting the class for majority voting is given by:

$$D(x) = \underset{c_i}{\operatorname{argmax}}\{v_i | i = 1, 2, \dots, k\}, where\ k\ is\ the\ number\ of\ classes.$$

3.2.5 Validation Protocols

Validation processes is important to deal with two major problems in pattern recognition. The first factor in such protocols is the model selection for data classification, in addition to related parameter estimation (e.g. k in k-NN classifier). The second problem is the performance estimation, which is measured by the true error rate of the data samples involved in the model testing. Performance measurements aim to estimate the true error rate of the whole "population".

Data splitting to training, validation, and testing sets is adapted to exploit the knowledge provided by the whole data in model training, validation, and estimating of the model performance. It is expected that training of a model result in an over-fitting problem, due to bias trend happen in optimizing the classifier hypothesis toward the training data. Consequently the training samples might be classified very well, while for an independent part of data the result might not sound comparable to the training samples. Cross validation via holdout, and leave one out approaches is used to overcome this problem.

The idea of holdout is to split the data in to k-folds; one fold will be used to evaluate the model and the remaining parts of the data to train the model. The popular number of folds used is k=10. If k=N where N is the number of the observations, the cross-validation technique called Leave One Out (LOO), which means that one sample will be leaved for testing the model while the remaining samples will be involved in training the model. Choosing the parameter k depends on the size of the data and the application under investigation. However small k results in less time consuming in model training, small variance of experiment performances, but the bias of the estimator will be large. For speaker independent applications, the data samples of an individual single speaker used to evaluate the model need is not involved in model training. Therefore, in this work Leave One Speaker Out (LOSO) approach is adopted for two of the database (Emo- Berlin, and the Kurdish), while in the case of Aibo FAU database we followed what suggested in the interspeech09 challenge by using the "Ohm" part of the data for model training and the "Mont" part to test the model.

3.3 Related works on feature Pre-processing in SER

In the literature feature selection is a frequently adopted process in emotion recognition, while only few studies are found to deal with meta-feature selection. This may reflect the ease with which analysis and interpretation of each individual feature can be conducted and the outcome can be related directly to the parameters of speech in its original space (Batliner , et al., 2011). For example, Vogt et al. (Vogt & Andre, 2005) adopted a Correlation based Feature Selection (CFS), using the best first search approach to select 90-160 from 1280 features. The tool used for feature selection was the publically available data mining software Weka. The study classifies the emotions by Naïve Bayes classifier, benefitting from the consistency of a de-correlated data produced by CFS and Naïve Bayes classifier, the simplicity in computation, and the satisfactory performance when the data has unbalanced samples per classes. When they applied on acted and non-acted datasets, they found that the selected features for acted dataset are mostly not overlapped with those selected from non-acted dataset. This highlights the difficulty of determining a specific limited set of feature for SER.

Pérez-Espinosa (Pérez, et al., 2012) proposed another filter-based forward feature selection approach that uses filter based correlation share and portion measurements to update the set of features. They concluded that the MFCC, LPC and cochleagrams feature groups are very important to estimate the level (High – Low) of the three Emotion Primitives (Valence, Activation, and Dominance). They built three models using SVM to classify emotional instances into the high\ low labels for each primitive. The primitive levels are used to train another SVM to assign the instance primitives levels to their basic emotions.

The above two filter-based approaches to feature selection may not be suitable for all classifiers. In the literature there are researchers who adopted the alternative "Wrapper" feature selection techniques for SER. Batliner et al. (Batliner , et al., 2011), adopted the Sequential Floating Feature Selection (SFFS) approach to select from a combination of three sets of features consisting of acoustics only, linguistics only, and both acoustic and linguistic based features. The number of selected features is fixed to 150 features (50 per for each of the three splits), using a cut-off criterion related to Receiver Operating Characteristic (ROC) curves. SVM classifier is used in error measuring as criterion for the SFFS. This approach, denoted by SVM-SFFS, was first used by Schuller et al. (Schuller, et al., 2005a), to select 75 features from 276 speech features and tested the performance of the resulting scheme on the Emo-DB database using 10-fold speaker dependent cross validation. They compared the accuracy achieved by 9 classifiers found that SVM outperformed all of other schemes with an accuracy of 87.5%.

Lee et al. (Lee , et al., 2011) proposed a binary-based feature selection using Bayesian Logistic Regression to benefit from diversity of relevant features for each pair of classes. The process resulted in 40-60 features per fold and pair of classes, and reported an accuracy of 41.57% Un-weighted Average Recall (UAR) for FAU-Aibo database based on the interspeech2009 suggested protocol for data splitting. Note that this dataset with the same protocol is adopted in all our experiments conducted in this thesis.

A comparison between features and meta-features is rarely adopted especially in terms of classification performance and complexity. This could be due to the capability of analysing the untransformed feature space using feature selection methods, while meta-features somehow fuses multiple features in a way that the contribution of each single participating feature cannot be determined easily. However, recently, there have been some promising efforts in finding various approaches to interpret meta-features. For instance Simmons et al. (Simmons, et al., 2015) propose a hybrid approach that uses a mutual information-based statistics to have a biologically meaningful interpretation for the PCA output. Meta–features compensate for the disadvantages of feature selection methods may end up with features that are correlated. For this reason in (Batliner, et al., 2008b) PCA was used to de-correlate the feature space and a cut-off based on the eigenvalues is used to select a constant number of features.

Dimension reduction of SER related feature space by transforming into a new space (meta-feature space) is adopted in (Zhang & Zhao , 2013) using six linear and non-linear methods for dimension reduction namely PCA, LDA, locally linear embedding (LLE), isometric mapping (Isomap), supervised locally linear embedding (SLLE), and a proposed modification of SLLE namely Modified SLLE (MSLLE). The study extracted 25 meta-features and applied to two databases, one of which used in this study (Emo-Berlin database). The best highest obtained accuracy rate was 78.56%, using MSLLE and SVM classifier. We are not aware of any study, which has adopted transforming more than 1k features, to extract meta-features. In chapter 5, we shall investigate the use of meta-features, which are selected from thousands of features and compared to the feature selection methods.

3.4 Related Works to emotion classification

Classification techniques used for emotion recognition, in the early stages of SER studies, include the most popular ones like k-NN and LDC (Petrushin, 1999;Long Pao , et al., 2005). These techniques are reported to be successful for non-prompted data (Lee & Narayanan, 2005). Later on, and as the number of the investigated features increased rapidly, these classifiers were found to be incapable of achieving similar level of success. For instance k-NN is not affective in high dimensional feature space, because the distance-based distinction after 20 dimensions decreases until there is no difference between the distances in very high dimension space (Beyer , et al., 1999). LDC also faces serious challenges to model multivariate Gaussian distribution in a high dimensional space, due to the difficulty in estimating the data covariance, especially when the number of dimension is significantly larger than the number of observations (Bai, 2011). However the LDC is more suitable for dealing with un-correlated features as long as the density of samples is sufficiently large. Transforming high dimensional data into a PC space facilitates the use of LDC. In chapter six of this thesis an investigation of this topic will be shown.

The state of the art studies have frequently used the SVM classifier, and reported it as a successful technique for high dimensional space. The SVM is an extension of the LDC with new conditions based on maximizing of margins of the separator hyper-plane. The SVM has helped achieving a good accuracy in the high dimensional feature space in many SER studies (Batliner , et al., 2011;Hassan & Damper, 2012;Lee , et al., 2011). However, SVM being a binary classifier there are many challenges for its extension into a multiclass classification tool that need to be dealt with appropriately. Some researchers adopted a pairwise (1 vs. 1) approach (Schuller, et al., 2009b;Batliner , et al., 2011) to build a multi-class SVM classifier. In (Lee , et al., 2011) a hierarchical binary tree is adopted, that is based on a pre-knowledge of similarity between classes at each level. For example when applied to the 5-class FAU-Aibo database, the pair of classes in the binary tree was (Angry/Emphatic vs. Positive, Angry vs. Emphatic, Angry vs. Neutral/Rest, Emphatic vs. Neutral/Rest, Positive vs. Neutral/Rest, Neutral vs. Rest). Hassan et al. (Hassan & Damper, 2012) proposed an automatic approach that uses Non-Metric Multi-Dimensional Scaling (NMDS) of the confusion matrices to organize the hierarchical binary tree, and improvement of the accuracy rate over various multiclass SVM techniques is reported for speaker independent SER. Their binary tree produced for the FAU- Aibo dataset is shown in (Figure 3-6).

Ensemble classification was also used for SER, for instance Schuller et al. (Schuller et al. 2005b) used Bagging and boosting ensembles showing comparable accuracy to what are achieved by classifiers like k-NN and SVM.

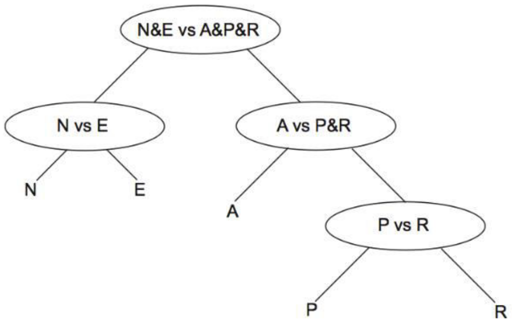

Figure 3-6: 3DEC scheme for five-class Aibo-Ohm database. Key: N-neutral; A-Angry; E-emphatic; P-positive; R-rest.

3.5 Summary

In this chapter, we attempted to cover theoretical background on the techniques of Computational SER in a sufficiently informative manner. This covered data pre-processing which include feature and meta-feature selection; and classification techniques, (SVM, ANN, LDC and Ensembles). Integrated into these discussions we had a literature review on both topics in SER area and related works. In the next chapter we initiate our investigations on SER feature/meta-feature selection and introduce new suggested set of features extracted from the LP-residual signal.

Chapter Four
Excitation source features for SER

Although emotions seem to be influenced by a wide range of speech signal parameters\features see (section 2.4); investigations of emotion related features are ongoing research area (Schuller, et al., 2009b; Pérez, et al., 2012). Features extracted from the LP-residual signal (also denoted by excitation source) are one of the feature sites, which characterize the glottal influence on speech signal and consequently reflect the embedded voice quality characteristics. Excitation source based features are not widely investigated in emotion recognition research. In the next two sections we shall provide motivation for the inclusion of excitation features and give detail description of different type of such features.

In the work presented in this chapter, we fuse the features extracted from the LP-residual with large set of feature from the original signal at the classification level using SVM and feed forward ANN. SVM is reported to model complex and real-world problems, and can control the complexity associated with high dimensional feature space (Anill & Robet, 2000). Additionally ANN is one of the most popular classification tools, which models the human process thought as an applicable algorithm (Heaton, 2008).

Our proposed model is tested on a newly created emotional database in Kurdish language, the Berlin emotional database (Burkhardt, et al., 2005), and the spontaneous FAU-Aibo database (Batliner, et al., 2008c; Steidl, 2009). A description of the datasets used in this study is presented in (Subsection 4.5.1.1).

This chapter is organized as follows: section (4.1) presents an introduction to the chapter, followed by a description of the set of features extracted from the LP-residual signal in section (4.2). In section (4.3) a description of a “brute force” set of features extracted from the original speech signal is presented, while section (4.4) explains the SER model methodology. Finally, experimental work and conclusion comes in sections (4.5) and (4.6).

4.1 Introduction

Speech signal is originally produced at the vocal cord, which is then subjected to filtration by the vocal tract and the lip, before being detected. The influence of the vocal tract on the original speech signal is modelled by a linear prediction analysis (LP-Analysis). Excitation source signal, known as the LP-residual, refers to the natural error that occur between the original signal and the predicted one. The influence of excitation source based features is mostly ignored in emotion recognition research. This could be due to: 1) the popularity of spectral features; 2) viewing the LP-residual as an error signal; or 3) Lack of knowledge about the higher order relations contained in the LP-residual (Koolagudi & Rao, 2012b). Practical experiments have shown the relevance of LP-residual signal to emotions. For instance, Chauhan et al. (Chauhan, et al., 2010) conducted a perceptual experiment using three sets of emotional utterances. The experiment used a total of 1200 utterances (15 sentences × 8 emotions × 1 artists × 10 sessions). The first set contains the original signals that carry information on both vocal tract and excitation source of the signal. The audio files of the LP-residual of the first set were computed to represent the second set. Finally, Exciting the LP coefficient of the signals in the first set using white random noise generates the last set. The last set contains the vocal tract effects on the signal. The Mean Opinion Score (MOS) of Average Emotion Recognition Rate (AERR) for original speech utterances by 20 listeners who were asked to recognize the emotions, was 60%, 45%, and 32% using the first, second, and third sets respectively. The human ability to recognize emotion from the source signal (45%) is a strong motivation to investigate the LP-residual signal.

Another motivation to use excitation source features is that the LP-residual contains information that cannot be predicted by the linear LP-Analysis due to the presence of non-linear relations between the speech signal samples. The non-linear information contained in the LP-residual captures the characteristic of the glottal closure instance region. The glottal vibration and variation (including the closure instances), captured from the LP-residual signal, is investigated in this study for their capacity to encapsulate discriminative information on different emotions.

In the next section, we shall formulate the LP-residual signal and describe different types of features that can be extracted from this signal. These features have been used for speaker recognition but not for SER (Soma, et al., 2012; Yessad & Amrouche , 2012). We propose to extract such features to complement the large set of features that are extracted from the original speech signal, and shall demonstrate the benefit from this addition to improve SER accuracy rates.

4.2 LP-residual signal proposed features

The LP-residual signal r of the speech signal is the difference between the original signal s and the predicted signal $\hat{s}$ obtained by LP-Analysis, i.e.:

$$r(n) = s(n) - \hat{s}(n) \qquad (4.1)$$

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \qquad (4.2)$$

Where a_k is the linear predictive coefficient that is determined by minimizing the error in the least square sense as described in (2.3.4), and p is the number of previous samples used in predicting the current one i.e. the number of linear prediction coefficients (p is set as 12 in this chapter). Figure 4-1 shows a speech signal, its LP-residual signal in addition to the spectral of the LP-residual signal.

As explained above, the LP-residual signal encapsulates information relevant to emotions in speech. Traditional SER schemes pay little attention to the LP-residual signal and instead focus on using different types of spectral and cepstral features (e.g. MFCC and LPCC) extracted from the original speech signal. Motivated by the discussion in the last section, we have investigated the addition of cepstral and wavelet octave features extracted from the LP-residual signal, to other traditionally used features extracted from the original speech signal for SER. Our LP-residual features are:

1. ***MFCC of the LP-residual signal***: For our investigations, 12 MFCCs is extracted from the LP-residual signal for each 20ms frame that is shifted with 10 ms (the first 12 coefficients is known to carry the most relevant information to speech). MFCC of the LP-residual signal is already used for speaker recognition (Soma, et al., 2012; Yessad & Amrouche , 2012) and

shows capability in detecting speaker related information. We shall propose it for SER to use it as another cepstrum of the signal *r*.

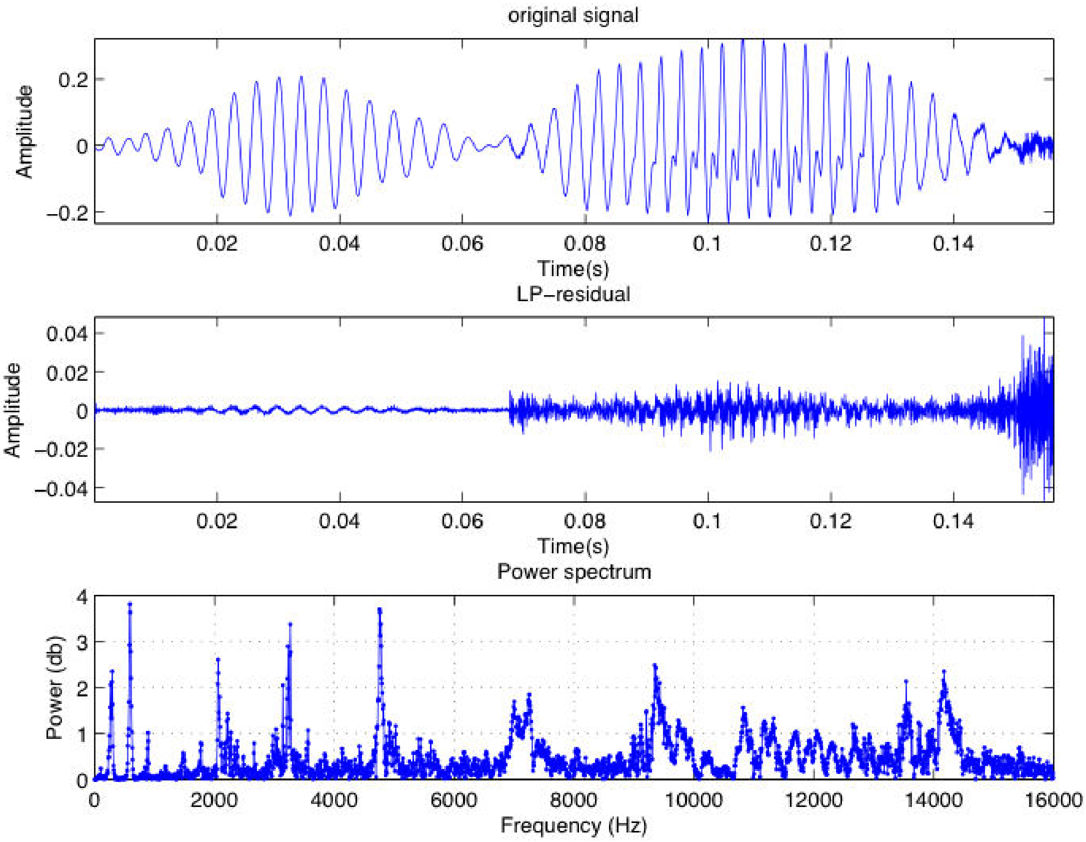

Figure 4-1: The spectrum of LP-residual signal of an emotional sentence.

2. ***LPCC of the LP-residual signal***: The second set of features includes 12 LPCC of the residual signal r are computed by applying LP-Analysis on the LP-residual signal to compute LP coefficients $\grave{\alpha}_k$ for each windowed frame (windows of 20ms with 10ms shift) followed by transforming them ($\grave{\alpha}_k$) into cepstral coefficients (see Figure 4.2). The $\grave{\alpha}_k$ is predicted using the same method of predicting a_k, such that:

$$\hat{r}(n) = -\sum_{k=1}^{p} \grave{\alpha}_k r(n-k) \tag{4.3}$$

where $\hat{r}$ is the LP-residual of the LP-residual signal

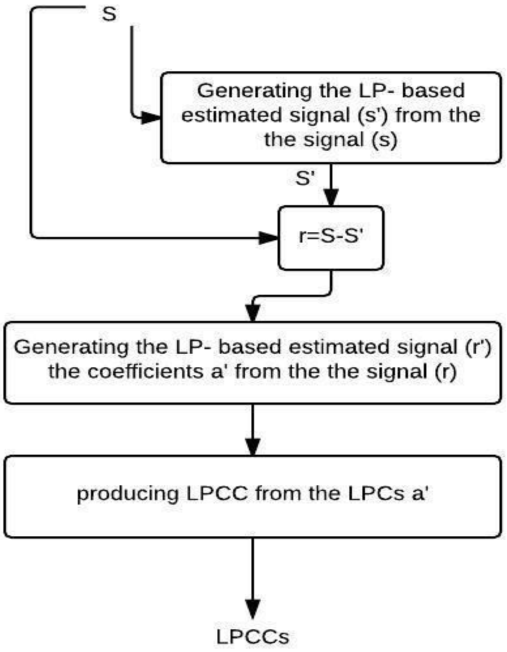
S
Generating the LP- based estimated signal (s') from the the signal (s)
S'
r=S-S'
Generating the LP- based estimated signal (r') the coefficients a' from the the signal (r)
producing LPCC from the LPCs a'
LPCCs

Figure 4-2: LPCC of LP-residual function

3. ***Wavelet octave coefficients of the LP-residual signal***: this feature refers to the Wavelet Octave Coefficients of the Residual (WOCOR) and has also been used for speaker recognition, (Zheng, et al., 2007). The motivation for using WOCOR, stems from the fact that, unlike Fourier transforms, wavelets are capable of characterizing the time-frequency properties in pitch pulses. To compute the WOCOR set of features, we first determine the voiced parts and estimate the pitch periods in the speech signal. Pitch period locations are used to apply Hamming windows with 2 pitch pulse periods. Wavelet transform is then applied to the windowed residual signal e_h (n) as shown below:

$$w(a, b) = \frac{1}{\sqrt{a}} \sum_{n} e_h\,(n)\ \psi^*(\frac{n-b}{a}) \qquad (4.4\ )$$

Where a={2^k k=1, 2, …, K} , b=1,2,…,N, and N is the windowed signal length. In this study we compute the WOCOR features as:

WOCOR={‖w(2^k,b)‖, b=1,2,…,N and k=1,2,…,6},

Where ‖.‖ is the Euclidian norm.

The above three types of features capture the cepstrum and wavelet band analysis properties of non-linear phenomena included in the LP-residual signal, which we use to demonstrate that the residual signal encapsulates important emotion related information. In total we obtain 54 LP-residual features made up of mean and standard deviation of 12 LPCC and 12 MFCC coefficients, as well as 6 WOCOR. Here after we shall denote this set of feature as Excitation Source features (ES).

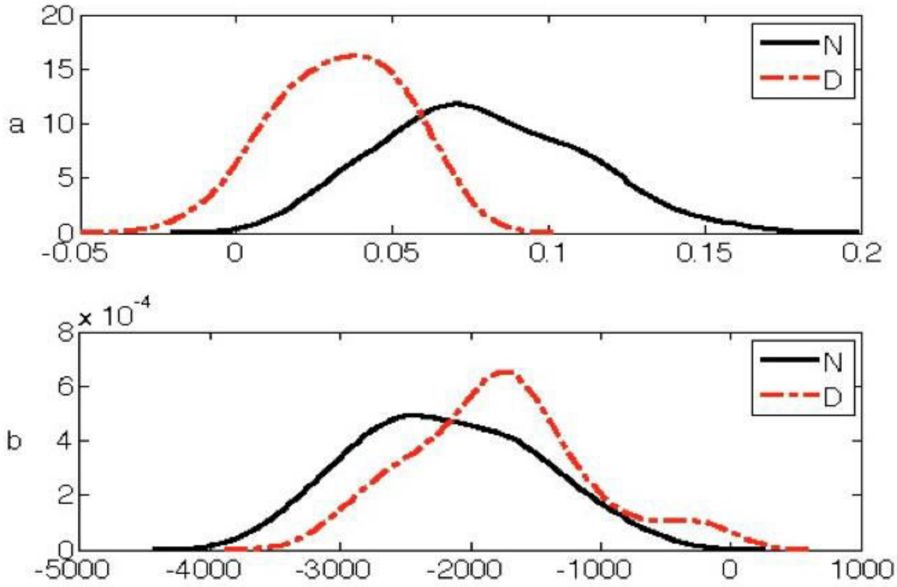

Figure 4-3: (a) the density distribution of 2nd MFCC's mean of the LP- residual for Neutral and disgust. (b) The density distribution of 2nd MFCC's mean of the original signal for Neutral and Disgust emotional samples from Emo-Berlin database.

It has been known that LPCC and MFCC of the original speech signal are among the most emotion discriminating features in speech (Pérez, et al., 2012). Here, we shall argue that the above ES features also have emotion discriminating power. To demonstrate this we shall use the whole neutral and disgust samples from the Emo-Berlin database. The density distribution of some of the proposed coefficients (here the second MFCC of the LP-residual) discriminate the neutral and disgust emotion better than the MFCC of the original speech signal (See Figure 4-3). This observation is another motivation in proposing the ES feature set.

Some of the existing SER models use a "brute force" set of Low Level Descriptor (LLD), to be described in next section. In the rest of this chapter we shall investigate the complementarity of our 54 ES features extracted from the LP-residual signal to "brute force" set of LLD feature.

4.3 Low Level Descriptors LLDs

The LLD baseline set consists of 6552 LLDs for emotion recognition are extracted using OpenEAR toolkit (Eyben, et al., 2009). These features include five groups of features, as listed in (Table 4-1), together with several statistical parameters as listed in (Table 4-2). For each descriptor group and each feature with their deltas, the toolkit computes about 39 statistical functional. Finally, for each sample dataset we extracted 6552 LLDs (OP features later on), which includes 52 parameters, their delta and the delta of their delta.

Table 4-1: (33) Low Level Descriptor (LLD) used in Acoustic analysis with Open Ear

Feature Group	
Raw Signal	Zero-crossing-rate
Signal Energy Pitch	Logarithmic fundamental frequency F_0 in Hz via cep- strum and autocorrelation (ACF). Exponentially smoothed F0 envelope.
Voice Quality	Probability of voicing (ACF(T_0)/ACF(0))$\frac{\text{ACF(T0)}}{\text{ACF(0)}}$)
Spectral	Energy in bands 0-250Hz, 0-650Hz, 250- 650Hz, 1-4kHz 25%, 50 %, 75%, 90% roll-off point, centroid, flux, and rel. pos. of spectrum max. and min.
Mel-spectrum	Band 1-26
Cepstral	MFCC 0-12

Table 4-2: (39) functionals and regressions coefficient applied to the LLD contour.

Functionals, etc.	#	Functionals, etc.	#
Respective rel. position of max./min. value	2	Quartiles and inter-quartile ranges	6
Range (max.-min.)	1	95 % and 98 % percentile	2
Max. and min. value - arithmetic mean	2	Std. deviation, variance, kurtosis, skewness	4
Arithmetic mean, quadratic mean	2	Centroid	1
Number of non-zero values	1	Zero-crossing rate	1
Geometric, and quadratic mean of non-zero values	2	# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks - overall arth. Mean	4
Mean of absolute values, mean of non-zero abs. values	2		
Quadratic regression coefficients and corresp. approximation error	5	Linear regression coefficients and corresp. approximation error	4

4.4 The Speech Emotion Recognition (SER) model

Having described the two sets of emotion-related features (ES and OP), we shall use these two sets to model the SER. In this section we describe the experiments conducted to determine the performance of the suggested ES features for emotion recognition. We aim to investigate the complementary characteristic of the ES features to the 6552 LLD set (OP) used in (Hassan & Damper, 2012; Schuller, et al., 2009b).

4.4.1 Multi-class SVM and ANN Models

The SER problem under investigation in this thesis is a multi-class problem. Here we use SVM and ANN on both of the ES and OP feature sets using One Versus One

(OVO) or pairwise model approaches. Basically the SVM is a binary classification but it has been adapted for multi-class problems, while the ANN is a multi-class classifier. Generally, there are different models for binary classifications like (One Versus All (OVA), OVO, Directed Acyclic Graph (DAG), and Unbalanced Decision Tree (UDT), (Hassan & Damper, 2010). Each of these different schemes can be depicted as a directed graph whose nodes are labelled by a splitting of the class sets. See Figure 4.4, for some examples of diagraphs for each of these models.

The OVO is the most common approach of building a multi-class SVM. Unlike DAG and UTD, the OVO approach is not influenced by the positions of classes on the classification tree, because all the possibilities of pairs of classes is taken into account. In fact it is a kind of ensemble algorithm in the sense that for each class it repeats (*c-1*) experiments (where *c* is the number of classes) and then the majority rule decision can compensate for misclassifying it by few models. However OVO model approach has the disadvantage of complexity growing when the number of class *c* is large, because the number of repeated models is $\mathrm{c}\,(\mathrm{c}-1)/2$ models, i.e. it is $O(c^2)$-complexity. However, in our case the number of classes being small (<8) which means that the number of repeated OVO models is not prohibiting.

Traditionally ANN designs a single system that has *c* nodes in the output layer, each node representing an output class. The major disadvantage of this single system of ANN, appear with higher dimensional feature and\or large number of classes involved. Higher dimensional feature with a multiple output node requires more complex NN architecture, and consequently it requires high training time (Ou & Murphey, 2007). However, similarly to the SVM, there are other popular ANN approaches such as the OVO, OVA, and P models Versus Q models (PVQ), where P is a subset of the whole classes of the problem and Q is the rest classes. Each of OVO, OVA, and PVQ are a collection of binary ANN models. Again the OVO approach produces *c-1* decisions for each individual class, and in total $n = \mathrm{c}\,(\mathrm{c}-1)/2$ decisions. The OVA, on the other hand, produces c decisions in total. The PVQ is a multi-level binary tree that at the first level, it builds a model for P classes against the remaining Q = c-P classes. At each branch of every subsequent level, it repeats the process with a smaller subset of the parent set of classes, until each branch is single class.

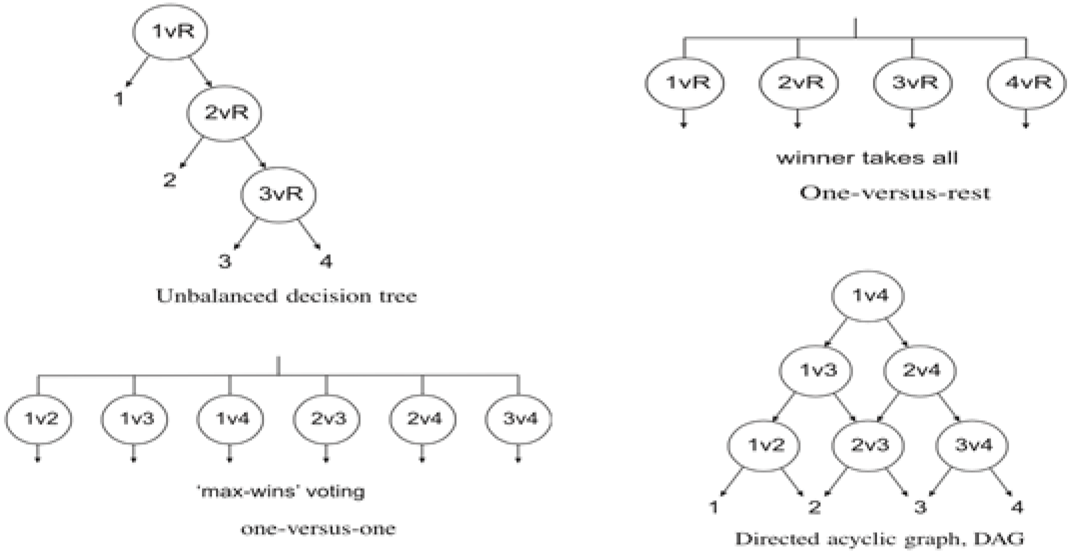

Figure 4-4: models for Multi-class SVM (Hassan & Damper, 2012).

In this chapter we adopt the OVO model for ANN. The OVO model is capable to reduce the complexity of each ANN system, since the number of classes is not large. Moreover, this would provide consistent way of fusing the ANN with the SVM.

4.4.2 A Fusion model

There are different ways of fusing the two OVO models of SVM and the ANN classifiers at the decision level. We are interested in determining how complementary ES features are to the OP features. We therefore conducted experiments by fusing OP and ES features using SVM and ANN classifiers, results in four possibilities of fusion. Given c classes, we obtain 2n decisions made by n SVMs in addition to n ANNs, $2n$ SVMs, or $2n$ ANNs for each experiment, where $n = c\ (c-1)/2$. The decisions made using OP and ES features for both classifiers are weighted by w_1 and w_2 that are validated using LOSO cross validation on the training set (i.e. 9-folds, 11 folds and 26 folds for Emo-Berlin dataset, Kurdish and Aibo datasets) and w_1+ w_2=1. In order to reduce the computation costs for the cross validation, the weights has been limited to {0.1, 0.2, …, 0.9}. Figure 4-5, shows the structure of an example of the adopted scheme.

To check the significance of the result throughout this thesis, we adopt one tailed binomial test, assuming that the results can be modelled as the discrete probability distribution of the number of the correct recognition trail out of the number of test samples (Hassan & Damper, 2013).

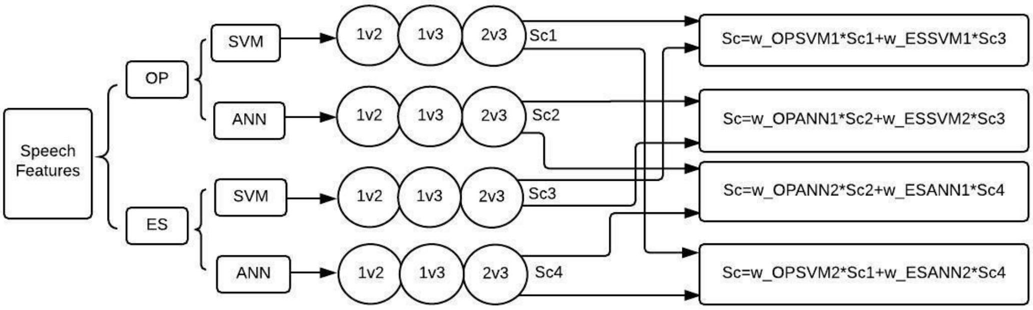
Speech Features
OP
ES
SVM
ANN
SVM
ANN
1v2
1v3
2v3
Sc1
Sc2
Sc3
Sc4
Sc=w_OPSVM1*Sc1+w_ESSVM1*Sc3
Sc=w_OPANN1*Sc2+w_ESSVM2*Sc3
Sc=w_OPANN2*Sc2+w_ESANN1*Sc4
Sc=w_OPSVM2*Sc1+w_ESANN2*Sc4

Figure 4-5: Fusing the SVM and ANN, for 3 classes, (Sc) is a score produced from the ratio of the number of correct decision to (c-1), where c is the number of classes (3 in this example). w_OPSVM and w_OPANN are the LOSO based validated weights for OP feature when applied to SVM and ANN classifier respectively. While, w_ESSVM and w_ESSVM are the LOSO based validated weights for ES feature when applied to SVM and ANN classifier respectively.

In these experiments, the SVM with linear kernel function and Sequential Minimal Optimization (SMO) method was used for classification. Although, non-linear kernel like Radial Basis Function (RBF) is more popular, it has been reported to have no significant advantage over the linear kernel when the number of features is higher than the number of instances (Hsu, et al., 2010). For the feed-forward ANN, the input layer contains 6552 neurons for the OP feature set, and 54 neurons for ES set. A single hidden Layer with 70 neurons for OP and 40 for ES are used. The number of neurons in the single hidden layer is usually chosen between the number of neurons of the input layer and that of output layer (Heaton, 2008). The output layer contains one neuron with hard-limit function i.e. the output of each ANN will be either 0 or 1. Furthermore the method used for weight updating is scaled conjugate gradient.

4.5 Experimental Work

To evaluate the performance of the fused SER model we adopted a speaker independent approach. We followed the cross validation protocol of LOSO for both Kurdish and Emo-Berlin database. For the Aibo database, we followed the advice of the interspeech09 challenge (see (Schuller & Batliner, 2009a)) and use the ‘Ohm’ part of the database for training and the ‘Mont’ part for testing the SER model. Before we present the results of our experiments, we next describe the datasets used in this study.

4.5.1 The experimental Datasets and related challenges

Emotional datasets vary in many aspects, like the spontaneity of the data, the number and age of participants, and the recording environment. In this section we first will describe three emotional databases used in this thesis, and then we shall discuss the problem of the balance between the number samples for different classes.

4.5.1.1 Datasets

In this subsection we describe the three emotional speech datasets used in this thesis, which are:

1. **The Kurdish emotional speech database**: This is a new database acquired using Kurdish language speakers. It includes 7 emotions (Anger, Happiness, Sad, Fear, Boredom, Surprise, and Neutral) acted by 6 male and 6 female actors. Each actor utters 10 Kurdish sentences on 4 different sessions for each emotion. Consequently, the dataset contains a total of 3360 recordings (i.e., 12 ×7 ×10 ×4 recordings). The speakers have been told to act these sentences and express them in their own style, by remembering/imagining a situation with the relevant emotion. The aim for building this dataset is to produce emotional speech samples that are expressed based on the actors ability and style of expressing emotion, and judged by native listener, in addition to extract some linguistic features. However linguistic features is not used in this thesis. The speakers have different acting experience (2 to 8 Years) and ages (19 to 36 years). The recording process has been done in a quiet room without restriction in the recording path. We selected 5 long and 5 short sentences to be spoken. The data files are recorded in wav format with 32 KHz sample rate. To determine the perceptibility of the emotions; 10 listeners participated in a subjective test. Average of correct labeling for the uttered sentences was 41%.
2. **Berlin emotional speech database**: The Berlin emotional speech database (also called Emo-Berlin) is an emotional speech database in German language. Ten professional native German actors (5 Male and 5 Female) were involved in recording 10 German sentences in 7 emotions. The considered emotions were Neutral, Anger, Happiness, Sad, Fear, Boredom, and Disgust. Some of the utterances were recorded in more than one session. Total of 535 utterances remained in the database after eliminating some unconvincing recordings based on a subjective test (Burkhardt, et al., 2005). The utterances are evaluated by 20 subjects as appropriate emotional sentences with

minimum score of 60%, assignable with minimum of 80%, and correct labeling accuracy rate of 84.3%.

3. **FAU-Aibo (spontaneous) database**: The FAU-Aibo database (Steidl, 2009; Batliner, et al., 2008a) was designed by recording children's sound, which is colored by different emotion, when they interact with Sony's pet robot Aibo. The children were led to believe that the robot is responding to their commands, whereas it was actually controlled by a human operator in a Wizard-Of-Oz manner. Sometimes the Aibo disobeyed the child's command, to lead them to different emotional reaction. The data were collected at two different schools identified by 'Ohm' and 'Mont'; the number of speakers was 26 and 25 respectively. Five experts labeled each word in the database independently, into 10 categories: angry, touchy, joyful, surprised, bored, helpless, motherese, reprimanding, emphatic, and 'other' for the remaining cases. The categories were mapped into four classes: anger, emphatic, neutral, and positive in addition to the fifth class for rest. The labels of the words were mapped to so-called 'turn' (i.e. utterances) and chunks using various heuristic method described by Steidl (Steidl, 2009). However, the dataset designer claims, "*Their 'turns' are similar to the units used in other studies. As they can consist of up to 53 words, they are not really optimal – we claim that our chunks are. By using different thresholds etc., the chunk size can be adapted to specific needs; the same way, different chunk sizes can be established for finding out how classifiers behave if faced with shorter or longer units. A pivotal characteristic of this solution is that our chunks are syntactically – and by that, semantically – well defined. This is a necessary prerequisite for higher linguistic (deep or shallow) processing in any end-to-end automatic dialogue system*" (Batliner, et al., 2008a). The number of chunks available in the dataset result in a total of 18216 speech samples. The Aibo corpus formed the focus of the Interspeech 2009 emotion challenge (Schuller, et al., 2011a).

4.5.1.2 Unbalanced number of samples to classes

The number of utterances per class in the FAU-Aibo databases is significantly unbalanced. For example, the number of Neutral samples in the Mont part of the dataset is 5377 samples, while there are only 215 samples that belong to the Positive emotional class. The unbalanced training influences most of the classification techniques significantly. Both SVM and ANN might bias to the classes with high number of samples. To overcome this issue, we follow the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is applied to the training set to increase the number of samples in the minority class (Chawla, et al., 2012). SMOTE generate samples on the line between any sample in the minority class and their k neighbours. Based on the desired amount of samples, *n* neighbours are chosen randomly, where $n<k$.

We also adopt UAR to measure the recognition performance accuracy of the FAU-Aibo dataset. UAR is defined as the accuracy per class averaged by the total number of classes. Note that the Weighted Average Recall (WAR) measures the total number of the correctly classified samples divided by the total number of the test set samples. UAR measurement is more realistic than the WAR for unbalanced data (Schuller, et al., 2011a). For instance in the Mont part of the FAU-Aibo database, 5377 out of 8257 samples belong to Neutral emotion (Here 5 emotional classes are involved). This means that if all the samples are classified as Neutral then the WAR will be equal to 65.1%, while in fact all the samples belong to the other 4 classes are classified incorrectly.

4.5.1.3 Speaker Normalization (SN)

In this thesis we applied Speaker Normalization on each individual speaker samples inspired by the work of Valsenko et al. (Vlasenko, et al., 2007). SN is realized as subtracting the mean of all samples that belong to one speaker, devided by the standard deviation of those samples. The aim of SN is to neutrlize the samples from speaker influence, thereby the emotion space is more adapted. While the emotion lables is not necessary in SN yet; its use will be applicable for speaker independent applications.

4.5.2 Results and discussion

To test the performance of the SER model designed in this chapter, we conducted four experiments. In the first one we fuse the OP feature classified by SVM and ES features classified by ANN. The second experiment fuses ES features classified by SVM with OP feature classified by ANN. The third experiment fuses both of OP and ES using SVM. And finally OP and ES features are fused using ANN. Recognition performance of each of ES and OP feature classified by SVM and ANN separately are also computed in this experiments. The results in Table 4-3 show the recognition performance of the suggested model applied to the three used datasets. A quick analysis of the results (in Table 4-3) reveals interesting patterns. The recognition performance using ES feature classified by SVM is 57%, 37.2%, and 36.7% for the Emo-Berlin, Kurdish and Aibo databases respectively. While in the case of all the three datasets, the same ES feature classified by ANN achieves less recognition performance (56.1%, 32.8%, and 35.6% with p=0.1, $p=6.2 \times 10^{-6}$and p=0.02 for the Emo-Berlin, Kurdish and Aibo databases respectively). In other words, the SVM has significant improvement of recognition accuracy over the ANN classifier for both Kurdish and Aibo datasets. However, the performance of the SER for the three datasets, when OP features are used follows a different pattern. The Emo-Berlin dataset achieve better recognition accuracy using SVM than ANN (p=0.02), while for the other two datasets the ANN outperforms the SVM with (p=0.04 and $p=7.6 \times 10^{-15}$ for Kurdish and Aibo dataset respectively). This quick analysis indicates that the adopted classifier alone does not influence the performance of SER; but there is an influence of the experimental database.

Table 4-3: SER accuracy rates, numbers in brackets are the fusion weights, Note that the number of categories a 7 for both Kurdish and Emo-Berlin dataset, while just 5 categories are available in Aibo dataset.

	Kurdish	Berlin	Aibo
ES (SVM)	37.2	57.0	36.7
ES (ANN)	32.8	56.1	35.6
OP (SVM)	43.2	87.4	39.3
OP (ANN)	44.4	84.7	44.9
OP(SVM)+ES(ANN)	44 (.8,.2)	89.0(.7,.3)	41.6(.7,.3)
OP(ANN)+ES(SVM)	**44.6(.8,.2)**	86.5(.7,.3)	45.1 (.7,.3)
OP(SVM)+ES(SVM)	43.8(.8,.2)	**89.2(.7,.3)**	41.1(.5,.5)
OP(ANN)+ES(ANN)	44.5(.8,.2)	85.4(.7,.3)	**45.5(.7,.3)**

A closer analysis of the results, reveal the following major observations:

1- The recognition performance of the SER for the Emo-Berlin datasets is much higher than the Kurdish and the FAU-Aibo datasets. We can attribute this to the facts that the Emo-Berlin dataset uses professional actors in recording the emotional data, unconvincing samples have been removed, and the database was designed for emotion synthesis and analysis not for emotion recognition.

2- The use of actors and the removal of large number of difficult samples from the Emo-Berlin dataset seem to result in having the trained samples cluster reasonably well around their classes which help separate them easily by linear classifiers such as the linear kernel used by the SVM. Note that the subjective test for Emo-Berlin dataset accuracy rate was 84.3%.

3- The OP feature vectors are of significantly higher dimension (6552) than the ES feature vectors (54). This may explain the reason for both classifiers achieve higher accuracy on the OP than on the ES features for all the databases.

4- For the Kurdish and FAU-Aibo datasets the classes samples seem to be more overlapped with regards to both feature sets, and this may explain to some extent the undesirable performance of the SER for both feature sets relative to the performance on Emo-Berlin dataset. Note that the subjective test that conducted on the Kurdish dataset result in 41% correct labelling of the speech samples. Accordingly, for the Kurdish database, our fusion scheme outperforms the subjective system and achieves 44.5%.

5- The linear SVM works on the original features with no dimension reduction, while the non-linear classifier (ANN with hidden layer) uses a supervised procedure to reduce the dimension of the OP (resp. the ES) set of features from 6552 to 70 (from 54 to 40). Usually, such procedures might help in reducing the effect of feature redundancy, which have an impact on accuracy rates. However, the impact on accuracy rate is somewhat dependent on the *ratio of the number of training samples used to the dimension*. These ratios for the Kurdish, the Emo-Berlin, and the Aibo datasets are 0.47, 0.07 and 1.52 respectively. Comparing the performance of the ANN and the SVM for the OP features seem to indicate that the nearer the ratio to 0 the better the performance of the SVM classifier, and as the ratio increases the ANN performance improves proportionately. This observation seems to be consistent with the conclusions of the work of Silverman, which investigated the difficulty in parameter estimation for kernels in high dimensional data (Silverman, 1986). We shall observe similar effects in chapter 5. However, this interpretation is not the only reason of high recognition accuracy for

Emo-Berlin dataset, but additionally the well-clustered samples for each class for the whole data help in avoiding the over fitting issue.

The main target of the experiments in this chapter is to investigate the complementarity capability of the ES feature to the large set of feature OP. For this purpose fusing both sets of features at the classification level is adopted. Fusing the proposed ES features with the OP improves the recognition accuracy from 44.4%, 87.4%, and 44.9% to 44.6%, 89.2%, and 45.5% for Kurdish, Emo-Berlin, and Aibo databases respectively. However, based on a one tailed Binomial test, the improvements over Emo-Berlin and Aibo datasets is significant with (p=0.03 and p=0.048 respectively). The stated weights reflect the amount of contribution of ES feature in improving the SER performance. The fusion results yield the following additional remarks:

1. In all of the four fusion schemes, fusing ES feature with OP features improves the SER accuracy, but they are significant on datasets. This is an indication of the complementarity between the ES features and the OP features.
2. The fusion weights (evaluated using LOSO cross validation on the training data) are almost similar regardless of the testing database or the classifier combination. For both EMO-Berlin and FAU-Aibo dataset optimal (OP, ES) fusion weight is (0.7, 0.3) which is comparable to the optimal weighting of (0.8, 0.2) the Kurdish dataset.

To be able to explain reasons for the very different patterns of accuracy observed across the 3 different databases, a closer look at the Confusion Matrix (CM) for each of the databases for all possible pairs of emotions (actual Vs predicted) could help identify the sources (classes of emotions) that are more responsible for false errors. The confusion matrix is a table that contains information the percentage of the actual classes (the rows) samples that the scheme can recognize for each of the predicted (columns) classes. Therefore the percentage of accurate classification for each class appears on the diagonal entry, and the confusion ratio between each pairs of classes appears at the other positions. Note that a 0 entry in the matrix means that the two corresponding classes are not expressed together, which might be addressing these pairs as opponent's classes, (see opponent theory of emotion, (Solomon , 1980)). The confusion matrices for the three datasets using the fusion model and ES feature are presented in tables (4-4 and 4-5) for the best performing fusion scheme as well as the best performing ES scheme. These confusion matrices convey the following information:

Table 4-4: Confusion matrix for Emo-Berlin, the Kurdish, and FAU-Aibo database for fusion scheme with best Average Recall.

		Anger	Happy	Neutral	Sad	Fear	Disgust	Bored
Emo-Berlin	Anger	92.1	6.3	0	0	1.6	0	0
	Happy	19.7	73.2	0	0	7.0	0	0
	Neutral	0	0	94.9	0	0	5.1	0
	Sad	0	0	0	93.5	0	6.5	0
	Fear	1.4	11.6	1.4	0	84.1	0	1.4
	Disgust	0	0	7.4	2.5	0	90.1	0
	Bored	0	2.2	2.2	0	4.3	2.2	89.1

		Anger	Happy	Neutral	Sad	Fear	Bored	Surprise
The Kurdish	Anger	73.8	13.8	4.8	2.3	0.6	3.5	1.3
	Happy	7.5	51.7	9.0	16.3	2.7	8.1	4.8
	Neutral	1.7	17.3	53.5	16.3	1.0	9.6	0.6
	Sad	4.2	24.8	18.1	29.2	6.9	13.3	3.5
	Fear	0.4	6.0	3.5	11.5	57.7	4.4	16.5
	Bored	5.6	26.9	15.4	22.5	7.1	18.1	4.4
	Surprise	5.4	18.8	5.0	6.7	35.2	4.6	24.4

		Anger	Neutral	Positive	Rest	Emphasise
FAU-Aibo	Anger	50.9	9.2	7.0	16.7	16.2
	Neutral	11.6	40.0	12.4	18.1	17.9
	Positive	2.8	21.9	51.6	18.1	5.6
	Rest	17.4	21.6	22.7	25.3	13.0
	Emphasise	20.1	20.6	2.8	11.7	44.8

1. For the fusion scheme, the CM matrix for the Berlin-Emo database has many 0's as a result of acting participants succeeding in suppressing all but the one they were asked to act. While the CM for the ES scheme has less 0's indicating to the low recognition performance of the ES scheme. To some extent, the confusion ratio of each individual emotion with the other emotions follow similar pattern of variation in both fusion and ES schemes for all the datasets.
2. The "Bored" and "Surprise" emotions are very badly recognized in the Kurdish dataset thereby affecting the overall accuracy. While in the Aibo dataset the "Rest" is the worst recognized categories, and extracted feature

vectors seem to confuse “Rest” with other emotions and only accurately recognise this emotion 25% of the time which close to the random selection of emotion (i.e. 20%). Note that Rest is defined in the FAU-Aibo dataset as the non-Neutral emotion that is not covered by the other categories (Batliner, et al., 2008a). The system presents this category as a scattered along the other categories, especially (Positive and Neutral). This might indicate somehow to the ambiguity of labeling the samples belong to this category, in the dataset design stage. It might also be an indication of having samples that includes a spectrum of other emotion information.

3. The Anger and Happy emotions (in both the Kurdish and Emo-Berlin dataset) are confused against each other. These two emotions are almost opponent in the valence dimension, but they have both high positions in the arousal dimension.

Table 4-5: Confusion matrix for Emo-Berlin, the Kurdish, and FAU-Aibo database for ES scheme with best Average Recall.

		Anger	Happy	Neutral	Sad	Fear	Disgust	Bored
Emo-Berlin	Anger	66.9	22.8	3.1	0	4.7	0	2.4
	Happy	42.3	38	7	0	8.5	1.4	2.8
	Neutral	6.3	2.5	60.8	5.1	0	24.1	1.3
	Sad	0	1.6	11.3	74.2	1.6	11.3	0
	Fear	14.5	10.1	1.4	8.7	58	2.9	4.3
	Disgust	1.2	4.9	28.4	12.3	2.5	48.1	2.5
	Bored	10.9	8.7	6.5	0	26.1	4.3	43.5

		Anger	Happy	Neutral	Sad	Fear	Bored	Surprise
The Kurdish	Anger	69	12.9	5.4	4.4	0.4	5.8	2.1
	Happy	14.8	31.3	15.4	17.3	3.5	9.4	8.3
	Neutral	6.7	14.6	46.7	16.9	1.3	10.6	3.3
	Sad	9.6	14.4	20.8	26.7	6	15	7.5
	Fear	2.3	5.6	2.9	11.3	52.7	7.1	18.1
	Bored	10.4	16.9	15	24	7.1	16.9	9.8
	Surprise	7.3	15.4	7.3	12.1	33.1	8.5	16.3

		Anger	Neutral	Positive	Rest	Emphasise
FAU-Aibo	Anger	39.8	10.3	15.5	11	23.4
	Neutral	19.2	27.9	20.5	13.9	18.6
	Positive	3.7	25.1	49.8	13.5	7.9
	Rest	20.5	17.6	26.4	19.6	15.9
	Emphasise	23	13.4	8.8	8.3	46.5

4.5.3 Comparison with state of the art studies

The ES features exploit much information of the LP-residual signal compared to some other proposed LP-residual features. For instance in (Koolagudi & Rao, 2012a) sixteen feature vectors each containing 40 samples of LP- residual around glottal closure is used as a source feature. The study achieves emotion recognition accuracy of 52.43% when applied to Emo-Berlin database. The nature of the feature extracted in (Koolagudi & Rao, 2012a) and the classifier used is quite different from what is used in this study, which make the comparison difficult. However it is obvious that the recognition accuracy achieved in this study is better using ES features in Emo-Berlin database (57%). Schuller et al. (Schuller, et al., 2009b) used the same OP feature baseline on Emo-Berlin and achieved SER accuracy of 85.6%. Another reported result by Vlasenko et al. (Vlasenko, et al., 2007) was 89.9%, when only 494 samples of the Berlin-Emo was involved, whereas our scheme achieved comparable accuracy of 89.2% but with the full set of 535 samples. The differences in the number of samples demonstrate that our results are as good as the state of the art, and validate our claim that the ES features are relevant to emotion recognition from speech. In the interspeech09 challenge many studies were made regarding the FAU-Aibo dataset (see Figure 4-6). The UAR achieved in majority voting among the whole proposed (44%), is significantly less than the achieved result in this chapter (45.5%), with (p=0.01<0.05, using one tailed Binomial respectively).

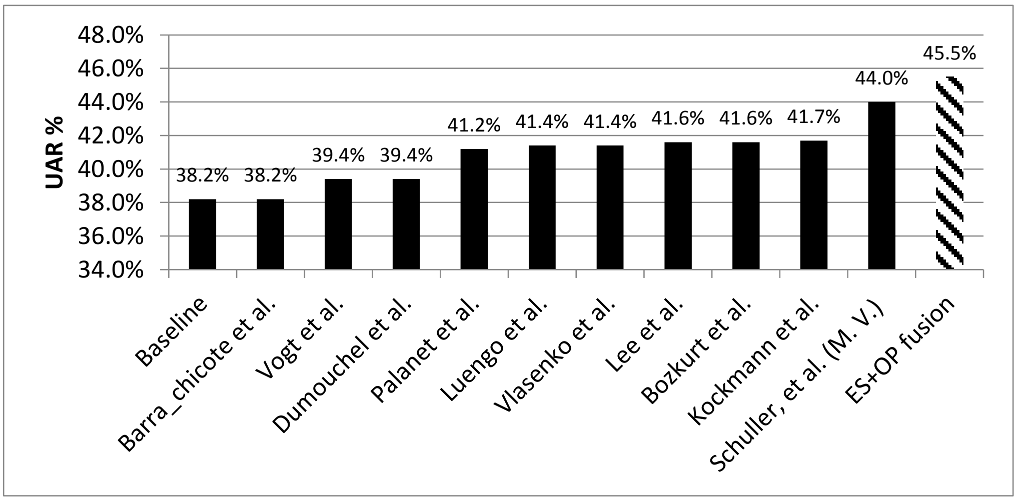

Figure 4-6: interspeech09 Challenge accuracy achievement including ten different works in addition to the majority voting of all of them (M.V.) (Schuller et al. 2011b). The last bar represents the result obtained in this chapter.

4.6 Conclusion

In this chapter we initiated a discussion on the need to use excitation source features, extracted from the LP-residual signal, for possible capacity to encapsulate discriminative information on different emotions. We introduced a relatively small set of ES features and investigated its complementarity to the high dimensional OP features that are extracted from the original speech signal. The ES features that we adopted here are the MFCC and LPCC of LP-residual signal as well as the WOCOR. The performance of ES-based SER, across various databases, using any of the two classifiers are sufficient to make the claim that the ES features alone do capture emotional related information, and the ES feature vectors (54 attributes) encapsulate sufficient emotion discriminating powers that would be complementary to the power of the traditional OP feature vectors (6552 attributes). In fact, the performance of the ES-only scheme for the Kurdish and Aibo databases are comparable to the performance of the OP-only scheme, while for the Berlin database the performance of ES-only scheme is satisfactory though much lower than that of the OP-only features. The work in this chapter also revealed the difficulties in separating different emotions unless individuals are well trained on expressing their emotion. In fact, these results are consistent with the difficulties mentioned in the psychology literature about the challenge of labelling emotions by individuals.

The high dimensionality feature space in these experiments is serious challenge in relation to system efficiency, hence the need for a dimension reduction step. Our experiments have confirmed the ability of ANN to recognize emotions even in the original high dimensional situation. This is due to the fact that the proposed ANN implicitly reduces dimension in the hidden layer. The ANN outperforms the SVM classifier on the Kurdish and FAU Aibo database. The fusions of ES and OP feature have performed as well as, if not better than, the state-of-the-art SER schemes. Based on these observations and conclusions, the next chapters will focus on the investigation of different dimension reduction schemes for SER.

Chapter Five

Features and Meta-Features selection for SER

The number of investigated features relevant to emotion from the speech signal has been increased in recent studies, which resulted in producing a high dimensional feature space. For instance in (Hassan & Damper, 2012) and (Schuller, et al., 2009b), 6552 features are involved in SER model training, and in (Batliner , et al., 2011) 3713 features followed by feature selection step are used for SER. However, the high dimensionality of any pattern recognition problem might produce two kinds of problems, model complexity and over-fitting.

This chapter investigates various solutions for dimensionality reduction, including sparse space-based feature selection, adaptive meta-feature selection techniques like PCA and non-adaptive ones like Random Projection (RP). Section (5.1) presents an introduction to the problem of high dimensionality of feature vectors and known solutions adopted in the literature. In section (5.2) we present a sparse-based feature selection scheme that is embedded in SVM, and we shall demonstrate experimentally the shortcomings of this approach. In (5.3) and (5.4), PCA and RP are described respectively as alternative solutions. In section (5.5), we shall analyse and demonstrate the performance of our proposed SER model using the PCA and RP projections for meta-Feature selection. Finally, the conclusion and a description of the next and final step in this thesis will be presented in section (5.6).

5.1 Introduction

Mathematical modelling of real life applications rely on providing an abstract representation of reality in terms of mathematical structures that encapsulate the main variables together with inter-relationships among the various entities/objects under investigations. This is a particularly tough challenge in applications where the main entities/objects/variable are not well understood or well defined. The case of SER is one such application that has shown to elude attempts to model mathematically, perhaps due to difficulties in having a well-defined understanding of human emotions from a psychology point of view. In such scenarios, often one builds models that involve a huge number of attributes that require dealing with a high dimensional vector space, even when one may not be able to ascertain that the variables (attributes) are uncorrelated or linearly independent. The various attributes that researchers have been considering in SER are normally assembled over a period of time with different refined considerations of the subject materials. This standard scientific refinement approach often leads to incremental improvement of the mathematical model performance but in many case attributes are introduced that are implicitly/nearly present in terms of a combination of others.

In general high dimensional vector spaces may provide a good representation of datasets obtained from the domain application. However, in real-life applications when number of samples are constrained, the question whether the representation is reasonably adequate or not, one needs to know if the number of samples is consistent with the number of dimensions. Theoretically the need of large number of samples in a high dimensional space is obvious, due to the fast exponentially growing of the high dimensional space volume. In a high dimensional space, experimental data obtained from pattern recognition tasks is usually spatially sparse (i.e. of low density), and significant statistics is difficult to estimate or to discover /interpret patterns or anomalies that may exist in the wider population. Reliable generalization of classification cannot be achieved unless the number of experimental samples is sufficiently large and clustered together in a reasonably separated manner. So, a single class largely populates each cluster and the distribution of samples is a good representation of the wider real population. Silverman investigated the difficulty in parameter estimation for kernels in high dimensional data (Silverman, 1986). The number of required samples as a function of the space dimension needed for kernel parameter estimation based on Silverman studies grows exponentially as shown in Table 5-1 (Webb, 2002). Therefore, as the dimension of the modelling space grows, more samples are needed to represent the data clusters properly. For instance, to preserve the density of the *n* samples data as it is in one dimension, about n^2 samples are required to represent the same data in two-dimensional space, and almost n^{d+1} samples when transformed to *d* dimensions.

The logical step to deal with such a high dimensional space is to reduce dimension especially when limited number of samples is available. We shall investigate dimension reduction as a mean of feature selection or meta-feature selection (feature selection in a transformed space).

Table 5-1: Number of samples needed to estimate the kernel function in high dimensions (Silverman, 1986; Webb, 2002)

Number of dimension	Number of samples
1	4
2	19
5	786
7	10700
10	842000

Traditional feature selection in multi-dimensional spaces exploiting correlations between various attributes (or other objective functions) is impractical and inefficient when dealing with high dimensional spaces. Moreover, the traditional feature selections are not only dependent on the training samples, but do not guarantee the optimality/uniqueness of the selected features (Schuller, et al., 2011b). For this reason some studies refrain from naming the individual selected features, due to the difficulties of generalizing them to different data and recognition models (Batliner , et al., 2011). For applications that are modelled by a high dimensional vector space, the alternative efficient approach is to use special data sparsity based feature selection.

In our particular application, we follow studies that use sparse SVM whereby the l_1 "norm" is adopted beside the Euclidian l_2 norm. The l_1 minimization penalty is adopted instead of the sparsity “norm” l_0. Minimization of l_0 is equivalent to minimizing the cardinality of the selected feature set, and is known to be an NP hard problem. The advantage of this approach over the traditional feature selection method is its low cost (Liu, et al., 2013). Although, the SER is modelled by a very high dimensional feature space, we are not aware of any research that exploits the sparse-penalty feature selection approaches. In the next section the use of the recent updates of these methods for SER is presented.

However, all kinds of feature selection result in loss of information that is provided by the whole set of features/attributes, unless the selected attributes can linearly generate (i.e. span) all the other discarded attributes. Loss of information, as a result of feature selection, can be avoided only if the non-selected attributes are redundant for all the samples. Assuming that the number n of samples is compatible with the requirements stated in the Table 5.1, then ideally the selected set of features must satisfy the following Lemma that links the set of n-dimensional vectors formed by the unselected attributes to those vectors formed by the selected one.

Let $X_{n \times d}$ be a set of n samples each is a d-dimensional feature vectors, expressed as an $n \times d$ matrix whose rows are the samples. Suppose that a feature selection procedure selects $X'_{n \times k}$, where k<<d.

***Lemma*:**

If each column vector $y \in \{X - X'\}$, *is a linear combination of all or some of the columns of* $X'_{n \times k}$ *(i.e.:* $y = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k), then\ X'$ *encapsulate information the whole set X.*

Proof:

The definition of linear combination guarantee that each vector in $X'_{n \times k}$, *influence the magnitude and directions of at least one vector that belong* $\{X - X'\}$.

This Lemma doesn't guarantee that the feature selection is optimal. For optimality, one needs to show that the number of columns cannot be further reduced, i.e. the columns of X' are linearly independent. These requirements make the task of feature selection an inefficient task in high dimensional spaces.

Consequently, another alternative approach that can have the advantage of not ignoring any individual attribute is the meta-feature selection approach. Here, a meta-feature refers to linear combinations of attributes. The adopted meta-feature selection approaches in this thesis are based on linearly transforming the original features from their high dimensional feature space into a lower dimensional subspace. Here, we must have the number of selected meta-features *c*<<d, where d is dimensions of the samples. The set of coefficients of a selected meta-feature form a column vector in what is known as the projection $d \times c$ matrix P, also referred to as a *dictionary*. The selected meta-features are computed by the matrix multiplication XP. Efficient dictionaries are sparse, and in the transformed space the similarities are somehow preserved (i.e. do not grow beyond acceptable bounds).

This chapter will include investigations of different versions of transformation dictionaries, including PCA produced from the training data. We introduce also another version of PCA named by Data Independent PCA (DIPCA) that is generated using independent data of the participated dataset in the SER model training. The aim of this approach is to benefit from the behaviour of features from various data that has the same characteristics. For instance we produce the PCA from an emotional dataset and use it to transform the feature of another emotional dataset in the SER model designing. However, the meta-feature selection methods cannot help in discovering emotion's relevant features.

In this chapter the use of PCA, different versions of independent PCA, and versions of Random dictionaries (Binary Random Projection (BRP), and the Toeplitz (T_n)), are presented for meta-features selection from the OP set of 6552 features.

5.2 Feature selection using doubly regularized SVM (DrSVM)

In a high dimensional space, feature selection under SVM is frequently investigated and various approaches are suggested (Wang , et al., 2006). Such SVM models solve the l_q minimization, for different distance functions q, defined as follows:

$$\min_{\boldsymbol{w},b,\xi} \ (\lambda \|\boldsymbol{w}\|_q^q + \sum_{i=1}^{n} \xi_i \ ) \tag{5.1}$$

$$s.t. \ y_i(x_i^T \boldsymbol{w} + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0, i = 1, \ldots, n$$

Where $\boldsymbol{w} = (w_1, \ldots, w_p)^T$ is the vector of regression coefficients, $\xi \ = (\xi_i, \ldots, \xi_n)^T$ is the vector of slack variables, n is number of instances in the training set, y_i represents the label of instance i, and $\lambda > 0$ is the regularization parameter.

Note that the l_0 is a measure of the sparsity of a vector $\boldsymbol{w}$, i.e.

$$\|\boldsymbol{w}\|_0 \ = card\{w_i | w_i \neq 0\} \tag{5.2}$$

Therefore, l_0 minimization is meant to select "fewest" attributes from the whole set of features. Unfortunately, the l_0 minimization is an NP-hard problem (Candes & Tao, 2005). However, the popular alternative way to solve this problem is an approximated solution using the l_1 minimization (Bach, et al., 2011; Carlos , et al., 2013).

The l_1 optimization is a convex problem that can be modelled using linear programing optimization methods. It is argued that the use of the l_1 norm within SVM (i.e. l_1-SVM) yields advantages over the traditional SVM especially in the presence of redundant noisy features. But when highly correlated features are present, the l_1-SVM tends to pick fewer features. In other words “when there are several highly correlated input variables in the data set, and they are all relevant to the output variable, the l_1-norm penalty tends to pick only one or few of them and shrinks the rest to 0” (Wang , et al., 2006). To overcome these disatvanteges (Wang , et al., 2006) sugessted the Doublly regulrized SVM (DrSVM) method that uses a mixture of l_1 and l_2 minimization. The aim is mixing these two minimization methods to perform feature selection with the same capability of l_1 minimization, while selecting/removing the highly correlated features as a consequent of the use of the l_2-norm minimization.

The DrSVM has the following optimization form:

$$\min_{\boldsymbol{w},b,\xi} \left(\frac{\lambda_2}{2}\|\boldsymbol{w}\|_2^2 + \lambda_1\|\boldsymbol{w}\|_1 \; \frac{1}{n}\sum\nolimits_{i=1}^{n} \xi_i \;\right) \tag{5.3}$$

$$s.t. \; y_i(x_i^T\boldsymbol{w} + b) \geq 1 - \xi_i, \; \xi_i \geq 0, i = 1, \ldots, n.$$

In this section we shall adopt the iterative algorithm proposed by Lui et al. (Liu, et al., 2013), which called ISVM3. This version of the DrSVM, achieves efficiency as a result of removing some computations from the iterative loop. The assumption of sparse density of the data in the feature space justifies the use of such classifiers.

5.2.1 The DrSMV classifier

In this section DrSVM-ISVM3 is adopted to test the capability of sparse SVM using OP features and conduct experiments to test its performance when used for SER. The model builds classifying machines for each pair of classes, $n = c(c-1)/2$ classifiers, where c is the number of classes. Majority voting on the n machines will produce the final decision.

The FAU-Aibo database contains an unbalanced number of samples per class; therefore, we follow the SMOTE procedure for balancing the training set of the FAU- Aibo corpus, and use UAR to measure the recognition accuracy.

In the case of Emo-Berlin and the Kurdish databases, we used the LOSO approach to define the test and train set for model accuracy measuring. With the FAU-Aibo database the Mont part of the data is defined as a test set and the Ohm part is exploited to train the model. The data has also been normalized based on Speaker Normalization procedure described in 4.5.1.3.

5.2.2 Experimental Results and discussion

The recognition accuracy of the SER designed model using DrSVM is presented in table 5-2, which shows the impact of the DrSVM on SER model accuracy compared to the traditional SVM using the OP feature. Basically, the DrSVM select almost half of the used features based on minimizing the cardinality of feature sets using the l_1 minimization. The table also includes SER accuracy when we simply use a Random Feature Selection (RFS), which select in random manner a number of features close to the selected number by DrSVM. This table shows that it is even outperformed by the RFS.

Here some observation and analysis of the table 5-2:

1. It is obvious that the DrSVM does not improve the SER accuracy over the traditional SVM when applied to the OP features; actually, it is either worse (on Kurdish and Emo-Berlin dataset) or comparable (on FAU-Aibo data set). Feature selection removes some features from the input to the SER model, which might result in missing "important" information.
2. DrSVM uses a non-adaptive feature selection, and therefore we extended our experiments to compare its performance with that of a genuine Random Feature Selection (RFS) that selects 3500 features (almost half of the OP features) randomly to be classified by the traditional SVM. The results show that RFS can perform better (when applied to the Kurdish and Emo-Berlin dataset) or comparable to the DrSVM (when applied to FAU-Aibo) dataset.
3. The pattern of accuracy observed when SVM, DrSVM and RFS applied to Aibo dataset is different from that achieved for the other two datasets; the recognition accuracy for FAU-Aibo dataset lies between (38.5% and 39.7%). This might be due to the high correlation between features of Aibo dataset as a result of the tight range of speaker ages (10-13 Years). Here the voice characteristic of both genders might not be present clearly. Consequently, unlike the Kurdish and Emo-Berlin datasets, the gender factor is not influencing the speech samples.

Table 5-2: SER model result using SVM fed by OP feature and the DrSVM

Datasets	Traditional SVM	DrSVM	RFS
Emo-Berlin	**87.4**	82.6	85.1
Kurdish	**43.2**	35.4	41.7
FAU-Aibo	**39.7**	39.4	38.5

For a clearer view about the recognition performance for each single individual emotion, we now present in Table 5-3 the confusion matrix of the DrSVM recognition performance.

Table 5-3: Confusion matrix of SER using DrSVM.

		Anger	Happy	Neutral	Sad	Fear	Disgust	Bored
Emo-Berlin	Anger	85.8	10.2	0	0	3.1	0	0.8
	Happy	18.3	69.0	2.8	0	7.0	0	2.8
	Neutral	0	0	91.1	0	1.3	7.6	0
	Sad	0	0	1.6	93.5	0	4.8	0
	Fear	8.7	14.5	1.4	4.3	69.6	1.4	0
	Disgust	0	0	9.9	3.7	1.2	85.2	0
	Bored	0	2.2	8.7	0	4.3	4.3	80.4

		Anger	Happy	Neutral	Sad	Fear	Bored	Surprise
The Kurdish	Anger	66.3	15.2	5	8.5	0.6	3.5	0.8
	Happy	9.8	41.7	12.5	13.3	5.2	14.2	3.3
	Neutral	1.7	16.7	44	20.4	5.4	10.2	1.7
	Sad	2.5	22.5	20.4	23.5	7.3	20.8	2.9
	Fear	1.7	9	7.3	16.9	37.7	11.5	16.0
	Bored	5	24.2	17.1	27.1	7.5	15.8	3.3
	Surprise	5.2	16.7	9.2	16.0	22.3	12.1	18.5

		Anger	Neutral	Positive	Rest	Emphasise
FAU-Aibo	Anger	54.5	16	10.5	7.7	11.3
	Neutral	19.3	37.9	16.3	11.5	14.9
	Positive	3.7	25.1	51.2	13	7
	Rest	23.3	30.8	20.1	17	8.8
	Emphasise	27.5	18.7	5.9	6	41.9

The pattern demonstrated in these CMs is almost similar to the patterns observed in the last chapter when ES and OP are fused using ANN and SVM. For instance the system confused Anger and Happy emotions in the Emo-Berlin dataset (the Happy emotion is not present in the Aibo dataset). These two emotions are almost opponent in the valence dimension, but they have both high positions in the arousal dimension.

Another example is that, bored and surprise in the Kurdish dataset and rest in the FAU-Aibo are highly confused with other emotions and hence are badly recognized. The presence of such widespread confusion significantly reduces the SER performance.

We conclude from the results in tables 5-2 and 5-3 that the feature selection by the DrSVM reduces dimension but does not outperform the SVM using OP feature and not even the completely random feature selection. Consequently, we shall investigate the use of meta-feature selection in transformed spaces. In the next two sections PCA and RP are described as a techniques for meta-feature selection.

For all experiments in the rest of the chapter, we shall continue to follow the same procedures/protocols but we will only use the SVM classifier.

5.3 Principal Component Analysis (PCA)

The PCA finds directions (components or base vectors) along which data samples have high variances. The eigenvalues of the samples covariance matrix, after subtracting the mean of the samples, measure the samples data variance along the directions of the corresponding eigenvectors. The eigenvectors generates the principal components of the model, and are arranged in the decreasing order of the correspondence eigenvalues. Computing the PCA is somewhat restricted to the number of samples available.

Meta-features selected from the PC space are reported to be difficult to interpret, in terms of the original attributes, because the transformed sample dimension scores is a linear combination of all origin samples scores (Schuller, et al., 2011b). However, recently, there have been some promising efforts in finding various approaches to interpret meta-features. For instance Simmons et al. (Simmons, et al., 2015) propose a hybrid approach that uses a mutual information-based statistics to have a meaningful interpretation of the PCA output in terms of the original features. These studies are encouraging further investigation on the use of meta-features for SER.

5.3.1 PCA limitation when number of samples is small

As mentioned previously PCA is computed through the solution of eigenvalue problem of the covariance matrix $C= AA^T$ of the data A, that are generated through their corresponding eigenvalues. In what follows we shall show that the non-zero eigenvalues of a matrix like C is equal to *n*, where *n* is the number of samples included in the data set A. We will start with showing that the non-zero eigenvalues of both AA^T and A^TA are the same.

Theorem 5.1 (Spectral Theorem): *Assume that A is a symmetric matrix (i.e.* $AA^T= A^TA$*), then A is orthogonally diagonalizable and has only real eigenvalues* (Leinfelder 1979).

But unfortunately the matrix A is not always symmetric and alternatively we can use the fact that any AA^T *and* A^TA is a symmetric matrix. By applying the spectral theorem to AA^T *and* A^TA we can reach the following proposition:

Proposition 5.1*: The matrices* AA^T *and* A^TA *share the same eigenvalues.*

Proof: Let v be the eigenvector of A^TA with eigenvalue $\lambda \neq 0$.

This means that: $(A^TA)v = \lambda v$.

Multiply both sides by A:

$$A\,(A^TA)\,v= \lambda Av.$$

$$AA^T(Av)= \lambda\,(Av).$$

This means that AA^T and A^TA have the same eigenvalues λ, but with different eigenvectors Av and v respectively.

Practically, if A is 500× 2 matrix then the non-zero eigenvalues of the 2 × 2 matrix A^TA are the same of the eigenvalues of a 500×500 matrix AA^T but with different multiplicities. An example of the datasets used in the SER models in this thesis is the Emo-Berlin dataset that contains about 450 samples in the training stage, while the number of dimensions used is 6552. In such a case, PCA cannot create more non-zero eigenvalues than the number of samples (450); consequently the number of samples will restrict the upper bound of the principal components. This disadvantage in addition to the fact that the PCA relies on the training data makes PCA not always practical.

5.3.2 Data Independent PCA (DIPCA)

The above stated limitations of PCA cannot be used to exclude the use of this well-known and widely used dimension reduction scheme. Hence it is desirable to have an efficient PCA-based scheme that is independent of a given training set, and yet retain PCA's important characteristics in that distances/similarities between projected vectors are not much different from their original vectors. For this purpose, we introduce the data independent version of PCA matrices, DIPCA, by constructing a PCA projection matrix trained on different datasets (Emotional speech data, Neutral speech data, and Non-speech data), to be used in transforming any new unrelated dataset. The main benefit from such an approach is that the projection transform is computationally independent from the dataset available for training the classification model. For SER, the complexity of model training is decreased using of DIPCA helps in investigating the usefulness of sharing information between different emotional and non-emotional speech datasets. In other words feature correlation characteristic of one emotional dataset has been investigated in adapting different emotional dataset.

In this chapter each of the PC projection matrix trained on each of three datasets (Kurdish, Emo-Berlin, FAU-Aibo) will be used as a DIPCA projection for the other two datasets.

5.3.3 Random projection (RP)

In recent years, research into different branches of mathematics has revealed that certain types of random projection matrices can provide a dimension reduction tools without greatly inflating distances (or distorting similarities) in the transformed domain. It has been shown that certain types of orthonormal random projections can preserve the Gaussian mixture clusters (Dasgupta, 1999). The high dimensional data projected into lower dimensional subspace by projecting the data onto a "random" matrix in which columns are not necessarily orthonormal, but rather they are independent and identically distributed (i.i.d).

Theoretically, RP has the advantages of being generated independently from the data, i.e. the projection matrix generation is always done offline from the training stage. In fact they are randomly generated based on specific distributions that keep the i.i.d characteristic of the columns of the dictionary. Consequently, it contributes in reducing dimensions for different datasets, independently of the classification.

In this chapter we shall use two different types of random matrices named as the Binary Random Projection (RBP), and Toeplitz (T_n). The elements of BRP are zero and ones distributed randomly with equal ratio. While the T$_n$ matrix (See 3.1.2) is normally designed as follows:

$$T_n = \begin{bmatrix} a_0 & a_1 & \cdots & a_n \\ a_n & a_0 & \cdots & a_{n-1} \\ \cdots & \cdots & \cdots & \cdots \\ a_1 & \cdots & a_n & a_0 \end{bmatrix}$$

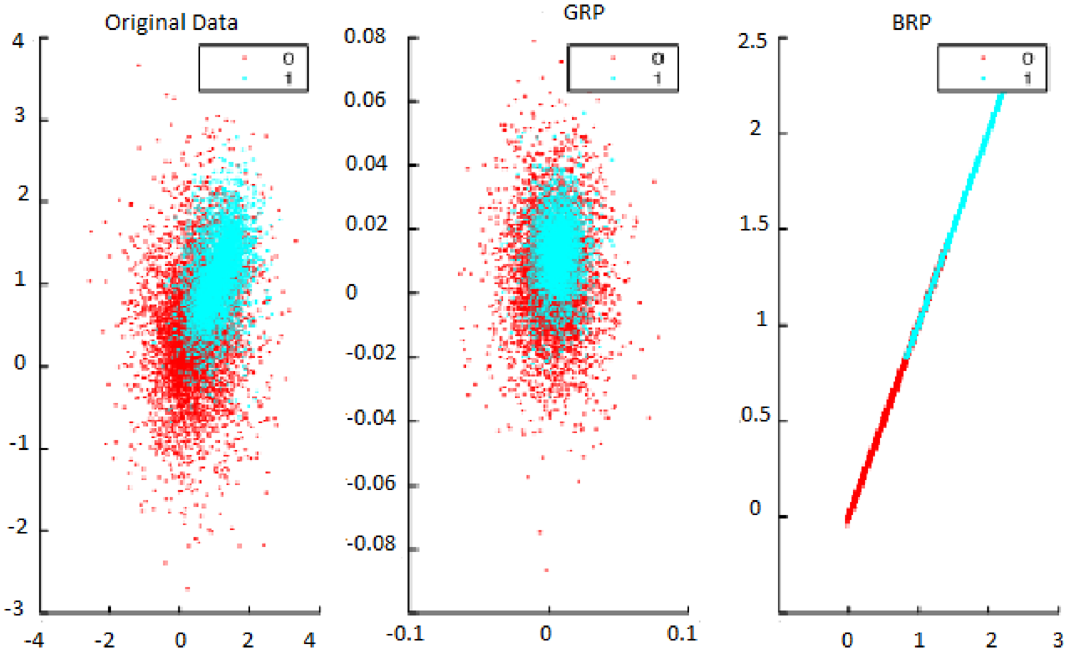

Figure 5-1: random data projected onto a Gaussian RP and Binary RP.

The projected data distances and/or similarities can be preserved with RP matrices that are i.i.d. and column element sums normalized to one. However in pattern recognition problems it is not necessary always to preserve distances. For instance non-linear kernel functions for SVM transform the data into new space in which their classes are linearly separated. In chapter three we showed that BRP matrices are not always orthogonal, but still contribute in data transforming to lower dimension, with reasonable preservation of clusters.

However, different random matrices behave in different ways based on the nature of the dictionary design. Next we present an experiment using a random generated data of two clusters with 500 samples in 6553 dimensions. The clusters are Gaussian distributed with their own mean and standard deviation that are chosen randomly. We project these data onto two spaces using Gaussian Random matrix and BR matrix. Figure 5-1 presents two of the 6553 dimensions that have been chosen randomly. Interestingly, we can observe how the distribution of the data is somehow preserved when projected on to the Gaussian matrix. While the projected data on to BR matrix loses their Gaussian distribution, but looks more correlated and separated. This behaviour of BR is encouraging to be used in SER model, using SVM classifier, in the sense that the projected data is better separated than the data in the original and the Gaussian space.

5.3.4 RP and DIPCA space adaptation

Real live SER applications are mostly speaker independent, i.e the objective is to recognise the emotion from a speech signal rather than recognising the person who is exhibiting the specific emotion. Therefore, speaker identities are not relevant to the training of SER. However, individual speakers included in speaker independent SER inevitably has an impact on the emotion classes clustering. Some researchers as in (Hassan & Damper, 2013; Zhang & Schuller , 2014) studied adaptation of emotional speech data, which contain different speakers, languages, or recorded under different conditions. The aim of the adaptation stage is to minimize the shift covariance between the training and test data. Hassan et al. (Hassan & Damper, 2013) propose an adaptation penalty, referred to as the Importance Weight, which is defined as:

$$\beta(x) = \frac{p_{te}(x)}{p_{tr}(x)},$$

where, $p_{te}(x), p_{tr}(x)$ are the multivariate probability distribution of the test and train feature vectors respectively. The estimated IW is used later as a penalty of the SVM classifier to adjust the hyper plane created by the SVM.

Zhang et al. (Zhang & Schuller , 2014) proposed another way for emotional data space adaptation that uses an auto-encoder based weight-adjusting procedure using the test and the train data to compensate for the expected shift that happened between the train and test data. The test data is used as an input to a one hidden layer De-noising Auto-encoder (DEA), which results in two matrices of weights W^1, W^2 for both the input and the hidden layer, in addition to a bias vector b. The output weights are used in another DEA that for the training data. The final output adjusted weights are used for both test and training data for adaptation purpose.

Inspired by these studies, we argue that the DIPCA would be useful in emotional data adaptation; in the sense that projecting emotional data onto the PCs of independently trained emotional data can adapt the test and train data to the same emotional based adapted space. Another possibility under investigation in this chapter is the use of RP for the same purpose. Although the RP does not depend on emotional data; but it has the ability to design an independent space that might be useful in any independent data adaptation.

5.3.5 RP and Kernel Functions

A kernel function transforms data into a higher dimensional space, such that the classes are more linearly separable than before. Kernel functions are exploited to find an easy environment for SVM in separating the classes linearly. For instance a data of two classes might be not linearly separable as shown in figure 5-2(a). A kernel function of form $z = x^2 + y^2$ transforms the data onto linearly separable regions as shown in Figure 5-2(b). However, in using kernel function for SVM the actual transformation is not necessary to be performed; but defining the inner product in the new space is adequate for its use.

In high dimensional space, however, kernel functions are not useful because they worsen the curse of dimension problem. Usually, in high dimensional space the class' samples are expected to be more separable from each other. The SVM is an efficient classifier for high dimensional data, but it still suffers from making a biased decision to the "non-adequate" training data samples compared to a high number of dimensions. Consequently, the issue of a biased decision toward the training set (over fitting), can be overcome by represent the samples in lower, rather than higher, dimensional space. Rahimi et al. (Rahimi & Recht , 2007) propose explicitly mapping the data to a low-dimensional Euclidean inner product space using a randomized feature map z: R^d → R^D so that the inner product between a pair of transformed points approximates their kernel evaluation.

Inspired by the experiment conducted in 5.3 that demonstrate that BR transforms/reduce dimensions while spreading out the samples, (see Figure 5-1); we propose the use of BR as a substitute for any kernel function with SVM. We shall demonstrate that the SVM hyper-plane can separate the classes post projection onto the BR space better than data in the original feature space and even when projected onto Gaussian Random (GR) space. However, we shall see later that this approach somehow increases the complexity of the SER system.

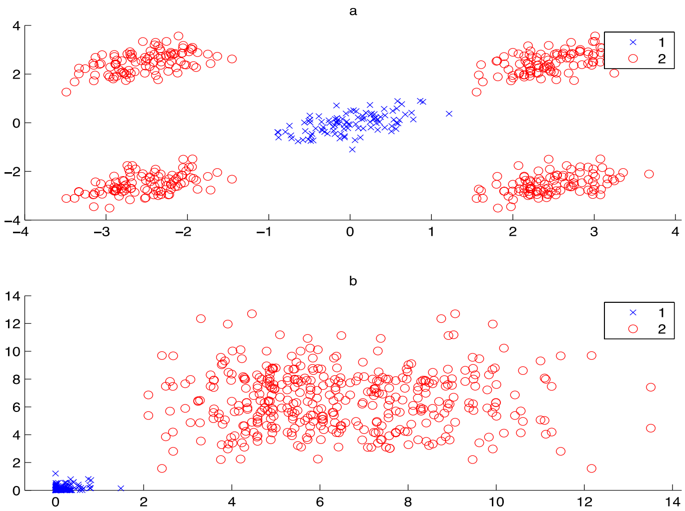

Figure 5-2: Example of data transforming to be linearly separated.

5.4 SER using meta-feature selection.

In this section we shall present the use of PCA, independent PCA, and different random projection matrices. The reduced dimension data will be fed into the SVM classifiers. The different version of PC projection matrices involved in the current experiment is generated from the training data, PCs of the other two emotion datasets, PCs of non-emotional speech (Neutral), and the PCs of non- speech data (image data). The image data is investigated to show how different environment space can contribute in creating proper projections for meta-feature space. The images were borrowed from a mammogram dataset, which are pre-processed by extracting features called Histogram of Oriented Gradients for Human Detection (HOG) feature (Dalal & Triggs , 2005). The random matrices used here are the BR matrix, and Toeplitz matrix. The BR matrix used in this study consists of *0s* and *1s,* with equal ratio. To be assured of the randomness effects on the recognition accuracy, 30 different versions of RPs are used in the emotion recognition model. The mean and the standard deviation of the results are presented. The experiments are conducted to test SER using a sufficiently large number of OP features.

5.4.1 Result and discussion

The experiments shown in figures 5-3, 5-4, and 5-5 present how data pre-processing with different transformation matrices contribute in the SER performance. The figures describe the changes of the SER recognition performance with different number of meta-features used in the classification model. However, these figures are presented to study the behaviour of these methods along different chosen dimensions.

The number of selected meta-features is one of the parameters, which need to be evaluated. Throughout this thesis, the classification parameters and model weights is evaluated by applying LOSO on the training set i.e., in this stage, 9-folds, 11-folds, and 26-fold cross validation are used for the Emo-Berlin, the Kurdish, and FAU Aibo databases, respectively. Tables 5-2, 5-3 and 5-4 present the recognition accuracy of each of the databases with the number of chosen dimensions in the evaluation stage.

From the figures 5-3, 5-4, and 5-5, and Tables 5-4a, 5-4b and 5-4c we can observe the following:

1. The recognition performance (UAR here) when using the PCA applied to the FAU-Aibo database decreased if more than 100 PCs out of 6552 PCs were selected (see figure 5-5), i.e. adding more meta-features lead to redundancy. The reason could be the close ages (10-13 years) of speakers involved in creating FAU-Aibo database, which lead to the absent of many variation factors (e.g. gender and age) and thereby increasing the correlation between original features, which mean that fewer meta-feature components suffice to adequately represent the training data. This pattern is not apparent in the other two databases. First of all these two datasets include recordings for adult male and female, and second there is a big difference between the ration of available samples in the two databases compared to the number of attributes: (1.52 for Aibo, 0.47 for the Kurdish, and 0.07 for Berlin).
2. The PCA projection generated by the FAU-Aibo dataset yields best SER performance when tested on all the 3 dataset. Some of the improvement could possibly be attributed to the fact that this projection was trained on the largest number of samples (See all individual tables).
3. The performance of the DIPCA projections, constructed from other emotional and neutral speech data, have a very similar pattern (the difference are marginal in either direction) to those obtained from training PCA on samples from the databases themselves. This is an indication of the similarity of the feature correlation of “any” emotional or neutral speech data.
4. The RP versions reach its highest recognition accuracy when selecting more than 500 meta-features. While the PCA needs (100-350 meta-features) to achieve its highest performance, which is attributed to fact that the first few PCs are the more “informative” dimensions. This is a consequence of RP randomness, which needs

larger number of representation of sample scores to be able to represent the whole meta-features.

5. When we trained PCA on the unrelated (to emotion) set of samples that come from the structured pattern recognition in mammogram images application, the corresponding PC_Hog projection achieves worse accuracy, although the performance improves when adding more components.
6. BR and T_n achieve the highest accuracies over PCA-based schemes, with (p=0.001, p=0.004 and p=0.0006 <0.05 for Emo-Berlin, Kurdish, and Aibo data sets respectively using Binomial test), but these non-PCA schemes consumes along time for training the classifier. While the PCA schemes consumes more time when a large number of samples in the training stage at projection generation stage.
7. However the time needed to test one sample varies from 6.1 ms to 17.3 ms, which depends on the number of features involved in the SER model, which is appropriate for real time application.

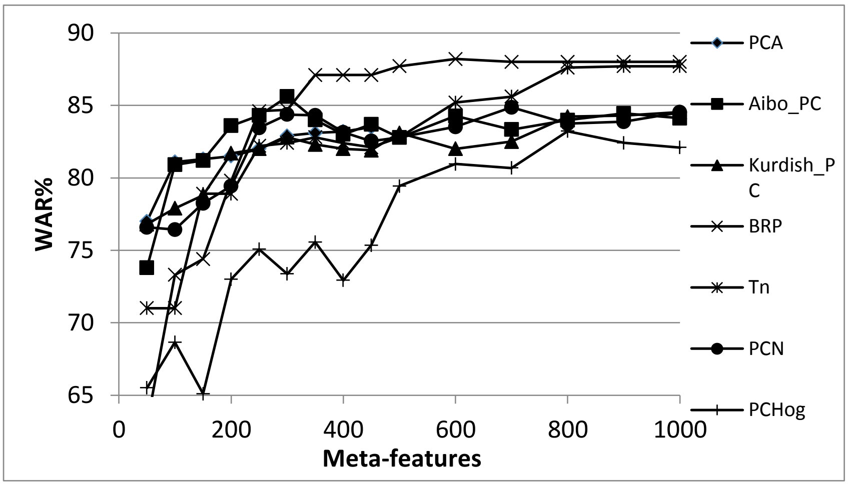

Figure 5-3: UAR for Emo-Berlin database, using different projection matrices

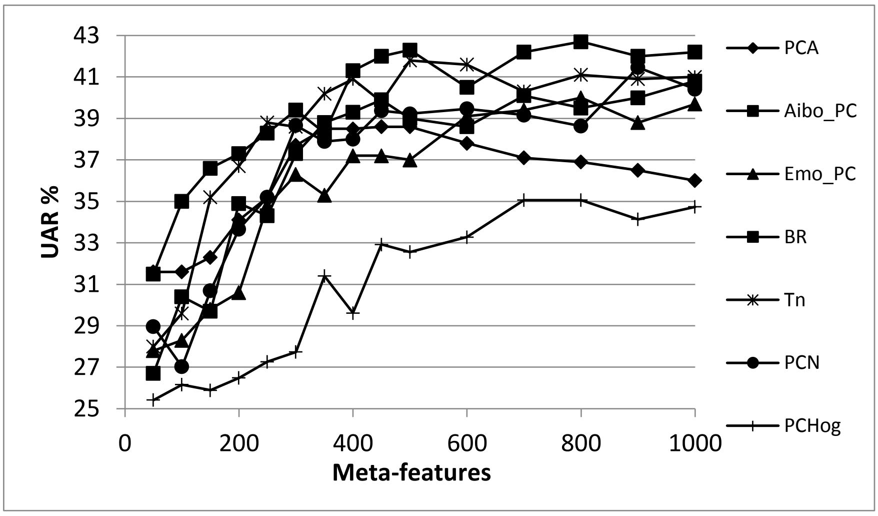

Figure 5-4: UAR for the Kurdish database, using different projection matrices

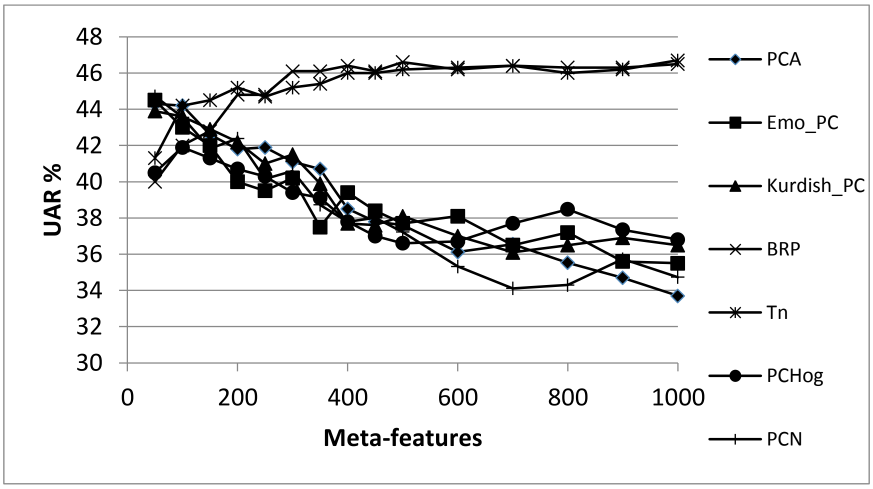

Figure 5-5: UAR for FAU-Aibo database, using different projection matrices

Table 5-4: Evaluated number of dimension and there relevant recognition accuracy (a: Emo-Berlin database. b: Kurdish database. c: FAU-Aibo database. LOSO in a. and b. while in c. the interspeech09 challenge standard is followed. The symbol (*) refers to the standard deviation of the experiments over 30 different version of RP matrix.
a: Emo-Berlin

	WAR	std(Sp)/Ac*	(Av/std) Dim	Ts_Time(ms)	Tr_Time (s)
BRP	**88.4**	4.5/0.97*	600/78.2	13.4	669.4
T_n	86.1	7.1/0.91*	770/48.3	13.6	886.9
PCA_Emo	83.4	7.3	315/66.9	10.1	85
PCA_Aibo	84.7	9.1	355/60	9.8	79.2
PCA_Kurdish	81.5	7.2	345/79.8	9.6	26.7
PCA_N	83.3	8.8	400/57.7	10.9	26.9
PCA_Hog	82.7	7.4	770/58	14.1	18

b: Kurdish dataset

	UAR	std(Sp)/Ac*	(Av/std) Dim	Ts_Time(ms)	Tr_Time (s)
RB	**42.2**	0.8*	710/69	17.5	3338.2
Tn	40.6	0.74*	620/45	16.3	3245.7
PCA_Emo	38.3	-	350/43	12.1	721.1
PCA_Aibo	39.4	-	400/56	13.6	761.3
PCA_Kurdish	38.2	-	350/48	12.7	827.9
PCA_N	38.5	-	410/51	14.1	721.3
PCA_Hog	35.2	-	780/43	23.1	185.4

c:FAU-Aibo database

	UAR	std (exp.)	Dim	Ts_Time(ms)	Tr_Time (s)
BRP	**46.5**	0.55*	700	16.9	2013.8
T_n	46.5	0.61*	700	17.3	2114.7
PCA_Aibo	44	0.34	110	6.1	2357.8
PCA_Emo	42.9	0.44	100	5.1	442.1
PCA_KURDISh	43.1	0.32	100	3.1	417
PCA_N	43.1	0.31	100	3.3	401
PCA_Hog	41.3	0.41	150	7.2	501.2

Finally, we investigated the adaptation role of transforming independent sets of emotional data (here testing and training data) into a random space. Figures 5-6, 5-7 and 5-8 show the UAR-in (the ability of the classifier to separate the training data), and UAR-out (the classification performance) for the datasets using PCA and BRP. Unlike the PCA, the BRP shows a stability of both UAR-in and UAR-out for the three databases and all dimensions up to 1000. Over-fitting appears early for PCA. Thus the BRP with SVM looks suitable for adapting spaces and avoiding over-fitting.

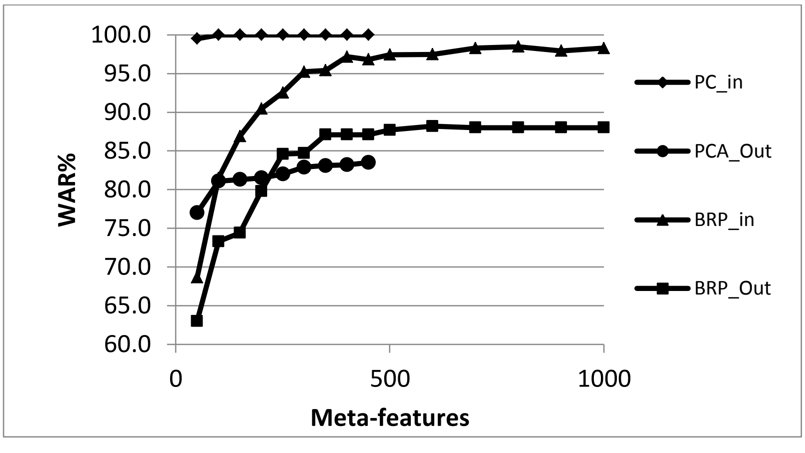

Figure 5-6: WAR_in &out using BRP and PCA applied on Emo-Berlin dataset.

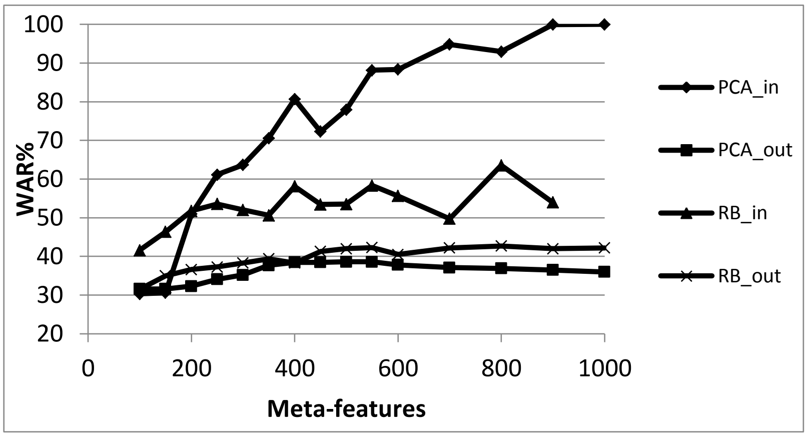

Figure 5-7: WAR_in &out using BRP and PCA applied on the Kurdish database

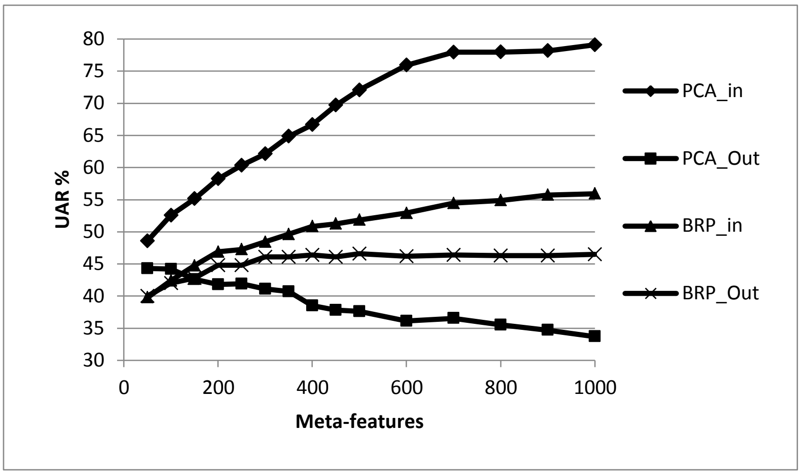

Figure 5-8: WAR_in &out using BRP and PCA applied on FAU Aibo dataset.

To further analyse of SER accuracy rates for individual emotions; the confusion matrices of the designed model using BRP+SVM is shown in table 5-5. The pattern is almost identical to the previously obtained patterns in using DrSVM (See table 5-3) and ES-OP fusion (Chapter 4).

5.4.2 Comparison to state of the art studies

It is necessary for the reliability of the obtained result in this chapter to have a comparison with the state of the art studies. We focus first on the schemes that use feature selection techniques applied on Emo-Berlin and Aibo datasets.

Schuller et al. (Schuller et al. 2005a), used SVM-SFFS by exploiting 75 features selected from 276 features and applied to the Emo-Berlin database using 10-fold Speaker Dependent (SD) cross validation, the suggested schemes achieves 87.5% recognition accuracy. It is obvious that the Speaker Independent (SI) approach that adopted in this thesis is more challenging than SD approach. Therefore, our achieved

SER performance (88.4% for SI) is significantly above what has been achieved by Schuller et al. (Schuller et al. 2005a).

Regarding the FAU-Aibo database, Lee et al. (Lee et al. 2011) propose a binary-based feature selection using Bayesian Logistic Regression. The process results in selecting 40-60 features per each fold and pair of classes and the achieved UAR was 41.6%. The protocol used was based on the interspeech2009 advices for data splitting, which is also adopted in this study. This result has been significantly outperformed by the use of BRP on Aibo dataset in this chapter (46.7%, with p=9.3×10^{-14}). The state of art results as well as the result achieved in chapters 4 and 5 is shown in Table 5-6.

Table 5-5: Confusion matrices of SER model using BRP+SVM

		Anger	Happy	Neutral	Sad	Fear	Disgust	Bored
Emo-Berlin	Anger	92.1	6.3	0	0	1.6	0	0
	Happy	16.9	74.6	1.4	0	5.6	0	1.4
	Neutral	0	0	93.7	0	1.3	5.1	0
	Sad	0	0	0	100	0	0	0
	Fear	1.4	7.2	2.9	0	88.4	0	0
	Disgust	0	0	7.4	6.2	0.0	85.2	1.2
	Bored	0	2.2	8.7	2.2	6.5	2.2	78.3
		Anger	Happy	Neutral	Sad	Fear	Bored	Surprise
The Kurdish	Anger	77.9	8.5	3.3	3.3	0.4	5.6	0.8
	Happy	9.4	36.3	9.8	17.9	4.4	16	6.3
	Neutral	4.2	9.8	49.8	20	4	10.4	1.9
	Sad	4.8	16.9	19.2	30.6	7.5	17.1	4
	Fear	0.8	3.5	3.1	13.1	52.3	5.4	21.7
	Bored	7.7	19	14.8	20.8	7.3	23.8	6.7
	Surprise	4.2	12.7	3.1	13.1	30.4	7.9	28.5
		Anger	Neutral	Positive	Rest	Emphasise		
FAU-Aibo	Anger	58.1	9.7	6.2	9	17		
	Neutral	11	42.2	15.2	14.5	17.1		
	Positive	2.8	22.3	60	11.2	3.7		
	Rest	19.6	23.1	25.8	20.5	11		
	Emphasise	19	17.6	4.6	8.9	49.9		

Table 5-6: A comparison of what achieved in chapter 4& 5 with the State-of-Art (SOA) methods.

Database	OP	BRP	ES+OP	SOA (FS)	SOA
Emo-Berlin	87.4	88.4	89.2	87.5 (Schuller et al. 2005a)	**89.9** (Vlasenko, Schuller & Wen 2007)
Kurdish	44.5	42.2	**44.6**	--	--
Aibo	39.7	**46.7**	45.5	41.6 (Lee et al. 2011)	44 (Schuller,et al., 2011b)

5.5 Conclusion

Recent SER schemes, including ours, begin with a pre-processing step that consists of extracting an initially high dimensional feature vectors and dimension reduction procedure that either opt for feature selection or meta-feature selection. We highlighted the practical difficulties that arise as a result of low density of available data relative to the high dimension of the feature vectors. This chapter was devoted to address the problem of feature selection and meta-feature selection as means of reducing redundancies that maybe a consequence of the original feature space being of a relatively high dimension. The background to the investigations in this chapter is extreme variation in the performance of exiting SER schemes, including ours, when applied to 3 different speech-based emotion databases. We opted for using the SVM classifier for its efficiency for high dimensional data, but we recognise that low density of available samples might lead to a bias decision towards the sparse training data. The use of DrSVM, within which feature selections are used to reduce dimension, has not led to improve the SER accuracy. Using tradition dimension reduction such as PCA, prior to applying SVM, was also used to overcome the problem without gaining much. However, the RP (BRP and T_n matrices in particular) prior to SVM, were effective in improving recognition. The RP matrices can avoid the problem of over-fitting and work like an adaptation process for the emotion feature space, and somehow compensate the need for the SVM kernel function. The over-fitting when using PCA appear with smaller number of components, especially in the Aibo dataset, in which the speakers voices features are more correlated due to the tight age duration.

We have shown in this chapter that “adequate” number of speech samples is useful to train PC projection matrix to be used later for classifying different dataset. The interesting result of using Neutral (non-emotional) speech data in training PC projections for different data involved in SER model, might reflect the neutrality of these recordings in the primitive directions like arousal and valence.

The need for balanced number of samples per classes is another issue that influences the performance of the SVM. For this reason it is useful for next step to know how this approach behaves with different classifiers that are not suffering from this shortcomings. In the next chapter the focus will classification level in SER including fusion of different models and different kind of ensembles.

Chapter Six
Classifications and Ensembles for SER

Having investigated the implication of feature selection and meta-feature, using a variety of dimension reduction procedures, on the SER performance across different types of emotional speech datasets we shall now focus on the effect of selected classifiers. In this chapter, we highlight some drawbacks of the classifiers used in this thesis, and investigate the impact of various alternative classifier schemes. In the last chapter SVM and its variant DrSVM were adopted for SER and considered as one of the effective and commonly adopted classifier in SER. However SVM suffers from inefficient time complexity, and more importantly for unbalanced samples per classes its decisions tend to be biased toward classes with larger numbers of samples. We shall investigate and develop alternative classification models that avoid the drawbacks observed in the models presented in chapter five (See section 6.1). In (6.2) we present the results of experiments conducted on an SER scheme that uses LDA with SVM. In (6.3), we describe, discuss and test the use of the LDC as an alternative classifier. Section (6.4) covers the performance of various forms of ensembles and multi-level classification that aim to improve the recognition accuracy of SER. We shall investigate the possibility of using the ensemble schemes for emotion ranking in (6.5). Finally the conclusion is given in (6.7).

6.1 SVM drawbacks

The classification scheme is among the most important factors affecting the performance of any pattern recognition model. The dimension of the feature space, the number of categories\classes, and\or number of samples might each has its role in choosing the proper classification technique.

Recent studies in SER, have frequently suggested SVM (Schuller, et al., 2009b; Batliner , et al., 2011; Hassan & Damper, 2012). It has already been tested in chapter 5 with various pre-processing techniques. SVM is an affective classifier for high dimensional data and has become well known in the last few years in many different applications, including emotion recognition in speech. But it has some serious drawbacks that affect the model complexity and sometimes the performance. These drawbacks are clearly observed as a consequent of the nature of the data chosen in this study in terms of being:

1. High dimensional features space with inconsistent number of samples, which normally results in over-fitting problem.
2. Imbalanced number of samples per classes in some of the datasets.
3. The multi-class nature of most of the SER applications.

In chapter 5, the issue of number of dimensions with the number of samples has been dealt with by meta-feature selection using PCA and RP. However the performance improvement of the PCA feature by the SVM model classifier was not as high as that for the RP projection. This might be because the selected meta-features using PCA are de-correlated and the density of samples in the reduced dimensional feature space is maximized; and thereby the SVM hyper-plane has less chance to separate the classes properly. While the BRP projection used in chapter 5 stretched the data such that the classes might be easier to be separated. However the use of BRP projection increases the complexity of SVM due to the difficulty in detecting support vectors in a random space. In this chapter we will investigate classification techniques consistent with the PCA like the Linear Discriminant Classifier (LDC), as an alternative for the SVM.

The second issue of imbalanced number of samples per classes is solved using SMOTE (Chawla, et al., 2012). SMOTE create artificial samples to increase the number of samples of the minority classes. Most of the classifier like SVM and ANN, require almost balanced number of samples per classes for proper performance, because the decision of these classifiers is based on individual samples. The LDC classifier, on the other hand, is based on modelling samples of each class using multivariate Gaussian distribution, which needs computing mean and variance of each class. Therefore, the imbalance of number of samples per classes is not a serious disadvantage as long as each class are represented by a sufficient number of samples to ensure reasonable estimations of the distributions parameters.

The third drawback is related to the difficulty of choosing a good multiclass model for the SVM, because SVM is a binary classification method. Multiclass SVM classification needs more than one SVM machine to make a final decision (Hassan &

Damper, 2012). Again this is not necessary for LDC because it is a multi-class classification technique.

In summary, LDC seems to provide a more appropriate alternative classifier for SER. However, LDC is not suitable for non-dense high dimensional feature space without a suitable dimension reduction technique like PCA, which produces a de-correlated set of meta-features. In the next section we shall present some experiments that also motivates choosing LDC in a de-correlated feature space.

One other classifier that is reported to be comparable in its performance to SVM for SER is the Random Forest (RF) (Schuller, et al., 2007). RF is one of the ensemble techniques that build many decision trees, each of which is fed by randomly selected subset of the data.

Inspired by the idea of ensembles approach to classification, we shall adopt different ensembles in this chapter. For example, by mimicking the RF approach using multiple LDC classifiers rather than decision trees.

The ensemble method could also be adopted, using various samples of instances, each feed independent model followed by a majority voting for having the final decision. The various samples could also be taken from the features, i.e. the feature set is divided into a randomly chosen subsets of equal size each feed an independent model.

In this chapter, we also propose yet another ensembles approach by choosing subsets of the meta-features, (using PCA in this chapter) to feed various models. We claim that subsets with different sizes of meta-feature present different representation of feature space, and will investigate its usefulness in SER in this chapter.

The different recognition accuracies of emotional datasets are also analysed in this chapter, based on how single emotion or emotion related state is available in the same speech sample. We refer to the usefulness of presenting the score of emotion classes for each tested sample instead of presenting the final predicted label. This might be helpful for some promising application like emotion content browsing from speech signal.

6.2 The Linear Discriminant Analysis (LDA)

In this section we shall present yet another motivation to use the LDC classifier for SER. Inspired by the fact that LDC is based on equivalent idea to the Fisher method of linear discriminating of classes (Fisher, 1936); we shall start by using LDA for dimension reduction, the SVM for classification, and follow the protocols used in chapter 5.

LDA is a supervised dimension reduction aims to construct a meta-feature space, in which the classes' samples are more separated. Fisher linear discriminant method aims to find directions, that minimize the within class variance and maximize the distances between the classes means. The objective of using the LDA in pattern recognition is manifested in reduced false rejection and false acceptance rates (i.e. increase the recall rate). In the last chapter, the meta-feature spaces were built using unsupervised methods such as PCA and RP. LDA can't be applied when the number of samples is less than the number of dimensions. To overcome this problem, the PCA is usually applied as a pre-processing step to transform the data into another subspace of a lower dimension, which is less than the number of samples. Figures 6-1, 6-2, and 6-3 show the recognition accuracy of this combined PCA-LDA meta-feature selection model, for our 3 experimental databases, using the SVM classifier. The results show increased recognition accuracy of the model compared to the use of PCA alone.

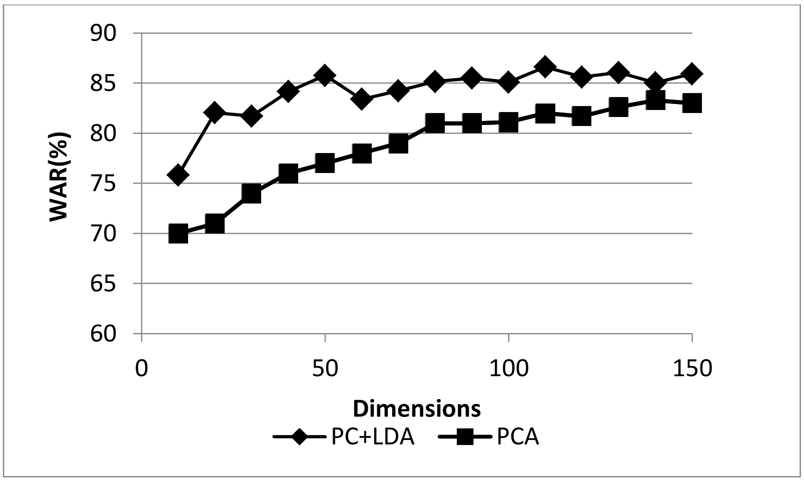

Figure 6-1: Recognition accuracy of Emo-Berlin using PCA pre-processing followed by LDA

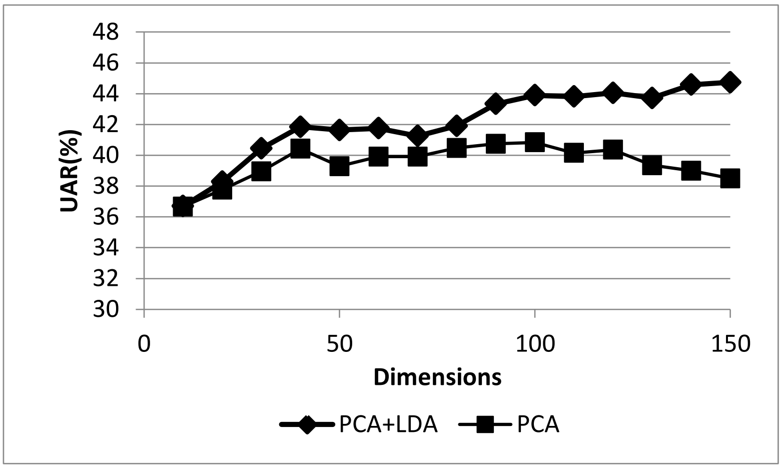

Figure 6-2: Recognition accuracy of The Kurdish using PCA pre-processing followed by LDA

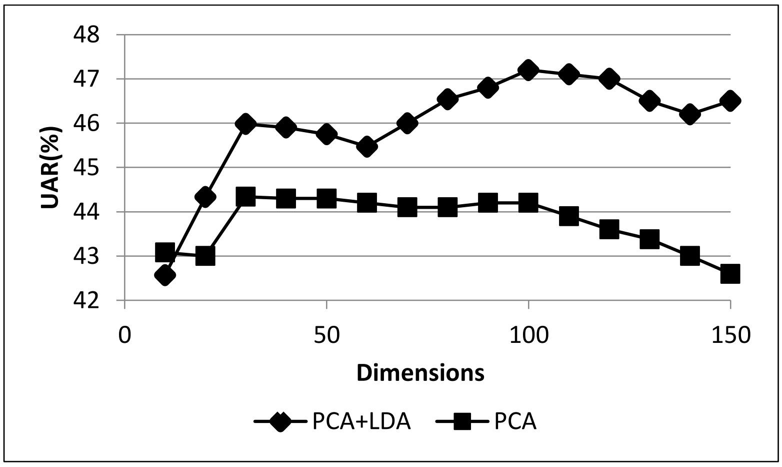

Figure 6-3: Recognition accuracy of FAU-Aibo database using PCA pre-processing followed by LDA.

In chapter 5, we presented the SER recognition accuracy achieved using PCA+SVM for the 3 experimental databases as 82.6%, 38.6 and 42.9% for Emo-Berlin, the Kurdish, and FAU Aibo datasets respectively. The above experiments show an average improvement of 4% - 6% in SER accuracy using PCA+LDA and the SVM classifier. In fact, this led to achieving 86.6%, 44.7% and 47.1% for Emo-Berlin, the Kurdish, and FAU Aibo datasets respectively.

Having established that LDA helps improving accuracy using the SVM, and the fact that the Fisher discriminating method underpin both the LDA as well as the LDC classifier then it is more natural to exploit this common factor to try and use LDC the classifier on the PCA instead of the SVM. This will be done in the next section.

6.3 SER using LDC Classifier

In this section we investigate the performance of a SER model based on using the OP features pre-processed by RP and PCA, with the LDC classifier. As mentioned earlier, the LDC assumes a multivariate normal distributed model for the samples of each class, which is usually approximated by using the covariance of the experimental samples. Again, all experiments follow the protocol that we described in chapter 5.

6.3.1 Results and discussion

The LDC relies on the assumption of multivariate Gaussian distribution of the clusters, and consequently works well with a de-correlated set of data (independent features). In the PCA, the eigenvectors (the principal components) that correspond to the most significant eigenvalues of the samples covariance matrix de-correlate the data and preserve the distribution of the projected samples much better than the RPs when using smaller number of dimensions. In tables 6-1, 6-2, and 6-3 the recognition accuracy achieved by the various meta-feature selection modes of the evaluated number of dimension for the three databases are presented.

Table 6-1: Meta-feature techniques using LDC for Emo- Berlin database

LDC-EMO	SVM (ch5)	WAR	std of RPs	Dim	Ts_Time(s)	Tr_Time (s)
BRP	88.4	80.7	0.13	115	0.068	23
Tn	86.1	75.3	0.08	155/43	0.08	24
PCA_Aibo	84.7	85.4	-	75/22	0.01	11
PCA_Emo	83.4	**90.7**	-	200/20	0.05	61
PCA_KURDISH	81.5	83.5	-	80/26	0.05	12

Table 6-2: Meta-feature techniques using LDC for Kurdish database

LDC-Kurdish	SVM (ch5)	WAR	std of RPs	Dim	Ts_Time(s)	Tr_Time (s)
BRP	42.2	39.7	0.1	245.8/14.4	0.05	11.5
Tn	40.6	38.6	0.12	229.2/33.4	0.04	23.7
PCA_Aibo	38.3	42.9	-	195.8/33.42	0.06	79.2
PCA_Emo	39.4	42.7	-	200/42.6	0.07	25.44
PCA_KURDISh	38.2	**43.5**	-	190.9/30.1	0.07	372

Table 6-3: Meta-feature techniques using LDC for FAU-Aibo database

LDC-Aibo	SVM (ch5)	UAR	std of RPs	Dim	Ts_Time(s)	Tr_Time (s)
BRP	**46.5**	43.1	0.4	500	0.06	45
Tn	46.5	42.6	0.31	400	0.04	54
PCA_Aibo	44	46.3	-	400	0.04	116.75
PCA_Emo	42.9	43.6	-	500	0.05	23
PCA_KURDISH	43.1	46.0	-	150	0.01	73

Analysis of the results in the above tables can be summarized by the following observations and conclusions:

1. In comparison with the previous result achieved when we used the SVM classifier; the recognition accuracy obtained by PCA+LDC is higher than PCA+SVM schemes when applied to all three datasets.
2. The LDC model is simple to be built, and more efficient than classifiers that include optimization process in separating classes. The training stage consumes negligible time though it increases as number of samples increase.
3. For each database, the best result of LDC is obtained when the PCA is trained on samples in the same database. This could be interpreted by the fact that PCA de-correlates the feature space; thereby the estimated multivariate Gaussian based parameters (e.g. mean and variance) are more accurate representation of the data than those obtained from other database samples. This also interprets the observation of that the PCA projections (PCA and DIPCA projections) outperform the RP projections in all the cases.
4. We are not aware of any SER work that use dimension reduction for more than 1K features, although it could be useful for simple classification techniques like the LDC.

The availability of different types of features/meta-features that are used in SER models and tested with classifiers is a motivation to investigate the use of ensemble classification in dealing with SER and the noted discrepancies in the pattern of accuracy achieved by the various schemes. In the next section we shall adopt different kinds of ensembles and test their performances for the 3 databases.

6.4 Ensemble classification

Ensemble classification consists of a set of models that are solving the same problem, with the aim of improving recognition accuracy. The models either use different classifiers trained by the same data (Boosting), or different versions of the data (e.g. differing in the selection samples/attributes) are used to train the same classifier (Bagging). The investigation made in this chapter is not interesting in using boosting approach, because the main aim of boosting classifier is to fit the as much as possible training samples. However, we observe that the data under investigation is not facing the problem of fitting the training samples, but rather the over-fitting is present. Consequently, the current investigations will focus on the Bagging approach to build ensembles for SER.

Choosing M random sets of samples with replacement, or M random sets of attributes facilitate the creation of variety versions of the same data. In both cases M models will be trained and final decisions are made according to a specific strategy that takes into account the decisions of all the models. Ensemble classification model trained by different random sets of samples/features is an attempt to train different models, in order to compensate for the incorrect decision of a classifier by other classifier(s). This might be helpful to avoid the drawbacks of the lack of "enough" samples for individual emotions. The ensemble model might be able to present the bigger picture of the SER model.

A common strategy is the majority-voting rule where all decisions are assumed to be of the same weight. However, the final decision can be based on the level of confidence/reliability (represented by a score computed from the model) of the various decisions. The score is a confidence measure how a single individual sample belongs to a class. In the case of LDC the test sample score is the posterior probability of each sample to belong to a class. The score gives more information about how a single sample belongs to a class in percentages (%). While the decision is scalar value refer to the class that the sample belongs to.

Here we shall propose the following three different ensembles designed for SER:

A. The first ensemble (Ens1 hereafter) consists of 5 models each of which trained using a set of random samples, chosen with replacement from the original data (66% of the data). While we are using LDC classifier, PCA is applied to each of the selected data and first 200, 150, and 400 PCs (the average number of PCs that achieved the highest recognition accuracy) are chosen for the Kurdish, Emo-Berlin, and Aibo datasets respectively. The majority-voting rule of the output class labels and the averaging of all of the classes' scores are computed to make the final decision.

B. For the second ensemble (Ens2), we trained 5 models using random sets of features (2000 out of 6552 features), followed by PCA dimension reduction to (200, 150, and 400) dimensions for the Kurdish, Emo-Berlin, and Aibo datasets respectively. As in Ens1, the majority-voting rule of the class labels and the averaging of all of the classes' scores computed to make the final decision.

   The aim of Ens2 is to investigate the ability of different feature subsets to influence emotion recognition, in addition to the possibility of modelling a more informative picture of the emotional feature space.

C. The third ensemble (Ens3), use three models, trained by data that are subsets of meta-features including the highest 50, 150, and 250 PCs. Although, the first 50 PCs are included in the first 150, which in turn is also included in the first 250 subset of feature, they are assumed to show different shapes of classes' clusters. These different meta-feature subset galleries increase the chance of all the samples to be represented using more transformed meta-features in the PC space. Unlike Ens1 and Ens2, the subsets of features here are not randomly selected but rather it is determinant way. Therefore, for the final decision fusion at the score level is adopted using these three subsets of meta-features. Figure 6-4 demonstrates the designed fusion model. In this model we are weighting the scores for each model. The weights are validated using LOSO approach applied to the training set, result in 9-folds, 11-folds and 26-folds validating models for Emo-Berlin, the Kurdish and the FAU-Aibo datasets respectively. Fusion of feature sets from different sites especially in biometrics is reported to improve the recognition accuracy.

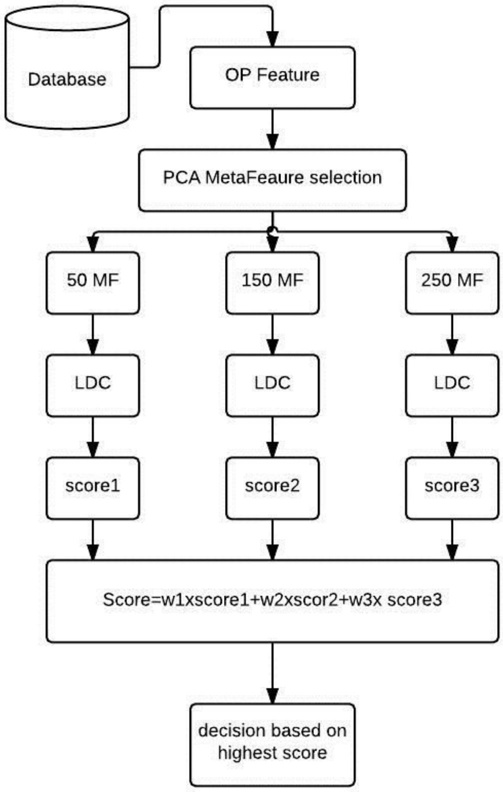

Figure 6-4: Meta-feature subsets using PCA fused with LDC classifier

6.4.1 Result and Discussion

In table 6-4 we present the results obtained using Ens1, Ens2 and Ens3. The highest accuracy achieved in the last section using PCA+LDC is shown in the first column in order to enable comparison with what is achieved by the 3 ensemble schemes. For Ens1 and Ens2, the results also include the standard deviation of the ten experiments in order measure the spread of accuracy for the different randomly chosen sample/attributes in the different experiments. The pattern of recognition accuracy, achieved by the ensemble models in table 6-4, convey the following observations:

1. The mean accuracy of Ens1 is marginally lower than that achieved by the PCA+LDC scheme, when applied to both Emo-Berlin and Aibo datasets. In fact taking into account the value of standard deviation over the 10 different repetitions, one can see that there would be random subsets of samples for which Ens1 has marginally better accuracy. While in the case of the Kurdish dataset Ens1 outperforms the PCA+LDC by more than 1.3% and taking into account the value of σ one can see that Ens1 outperform the PCA+LDC for almost all random subsets of samples (μ-5*σ > 43.5), i.e. the random data

subsets have more discriminating capability than the full set of samples after the application of PCA. This probably reflects the diversity of speakers experience in expressing emotions. Note that, when the mean of recognition accuracy of Ens1 is higher than the PCA+LDC baseline by more than 3σ, where σ is the standard deviation of 10 experiments.

Table 6-3: Recognition accuracy of Ens1, Ens2 and Ens3. MV refers to fusion at decision, and Sc to fusion at the Score level, while μ and σ refers to the average accuracy of 10 repetitions of experiments.

	PCA+LDC	Ens1		Ens2		Ens3
		MV(μ \ σ)	Sc(μ \ σ)	MV(μ \ σ)	Sc(μ \ σ)	
Emo-Berlin	90.7	90.2\0.35	90.4\0.41	90.9\0.48	90.9\0.46	**91.2**
Kurdish	43.5	44.8\0.17	44.9\0.25	**45**\0.33	44.8\0.24	44.9
Aibo	46.3	45.9\0.33	45.6\0.41	47.1\0.25	46.8\0.17	**47.3**

2. The Ens2 improves SER recognition accuracy above that of the PCA+LDC when applied to all the datasets. In other words, a collection of random subsets of OP features provides appropriate representation of emotion classes.
3. No significant difference is observed between both majority rule and score fusion decision for Ens1 and Ens2 schemes. The score based fusion is more efficient when the models is fed by deterministic data (as in Ens3), thereby weighting of the models can be adopted to highlight the useful ones in making the final decision.
4. Fusion using Ens3 outperforms the PCA+LDC scheme with significant accuracy ratio (45%, p=0.01 and 47.3%, p=0.02) for both Kurdish and Aibo datasets respectively. However in the case of Emo-Berlin dataset the 91.2% accuracy achieved by Ens3 is not significant at the 5% level.
5. In relation to the performance of the Ens3, we argue that the first 50, 150, and 250 PCs are different representations of the data; thereby weighting the models trained by those PC subsets improves the recognition accuracy.
6. We validated the weights using LOSO approach cross-validation. For each speaker in the test set we generated a specific set of weights based on the rest speakers' samples in the training set, i.e. the set of weights reflects how different sets of speakers' data could contribute in model weighting. The set of weights generated for Emo-Berlin and the Kurdish dataset are 10 sets and 12 sets respectively, because we apply 10-folds and 12-folds LOSO in testing

the model. In the case of the Aibo dataset, the weights are chosen for a single experiment. Table (6-5) shows the validated weights in Ens3 per each dataset.

In summary, the limited amount of experiments demonstrates the use of Ensemble approach to improve SER accuracy above the traditional deterministic approaches.

Table 6-4: Sets of validated weights per datasets in Ens3.

Emo-Berlin				The Kurdish				FAU-Aibo		
Sp.	50PC	150PC	250PC	Sp.	50PC	150PC	250PC	50PC	150PC	250PC
1	0.4	0.3	0.3	1	0.3	0.3	0.4	0.3	0.3	0.4
2	0.4	0.5	0.1	2	0.3	0.4	0.3			
3	0.5	0.1	0.4	3	0.5	0.1	0.4			
4	0.3	0.5	0.2	4	0.1	0.6	0.3			
5	0.5	0.3	0.2	5	0.1	0.8	0.1			
6	0.2	0.5	0.3	6	0.5	0.2	0.3			
7	0.5	0.3	0.2	7	0.2	0.4	0.4			
8	0.5	0.4	0.1	8	0.4	0.5	0.1			
9	0.4	0.4	0.2	9	0.3	0.4	0.3			
10	0.5	0.4	0.1	10	0.5	0.2	0.3			
				11	0.4	0.3	0.3			
				12	0.3	0.6	0.1			
Av.	0.42	0.37	0.21	Av.	0.32	0.4	0.275			
Std	0.1	0.13	0.1	Std	0.14	0.2	0.11			

6.4.2 Multilevel SER system

We now investigate a Multi-level Classifier (MC) approach to be built from the Ens3 confusion matrices, which are presented below in Table 6-6, for each dataset. This rather limited experimental work on MC classifier presented here are only a pilot study, inspired by the work of Hassan et al. (Hassan & Damper, 2012) who devised an interesting NMDS graphical tool to represent similarities between classes obtained from the confusion matrices. The bigger the confusion entry the more similar their class are. Hassan et al. used this graphical representation of the confusion matrix in building hierarchical classifier models whereby the more confused classes\emotions are combined in super classes to be separated from the other classes at the first level.

Given a confusion matrix, which may or may not be symmetric, with entries $c_{i,j}, i,j = 1,2,\ldots,C$ the k dimension NDMS is constructed by finding a set of k-dimensional vectors $x_1, x_2, \dots, x_C$, in which satisfies the following minimisation problem:

$$\min_{x_1,x_2,\dots,x_C} \sum (\|x_i - x_j\| - c_{i,j})^2$$

Here, $\|.\|$ might be a metric or non-metric measurement.

Table 6-5: Confusion matrix of the result obtained by Ens3.

		Anger	Happy	Neutral	Sad	Fear	Disgust	Bored
Emo-Berlin	Anger	94.5	3.9	0	0	1.6	0	0
	Happy	12.7	83.1	0	0	4.2	0	0
	Neutral	0	0	98.7	0	0	1.3	0
	Sad	0	0	0	96.8	0	3.2	0
	Fear	7.2	5.8	7.2	0	79.7	0	0
	Disgust	0	0	1.2	2.5	0	96.3	0
	Bored	0	4.3	2.2	0	4.3	2.2	87

		Anger	Happy	Neutral	Sad	Fear	Bored	Surprise
The Kurdish	Anger	70.2	12.7	4.4	1.7	0.4	8.5	2.1
	Happy	6.9	40.6	7.5	17.1	4	16	7.9
	Neutral	2.1	11.5	51.3	16	2.7	13.8	2.7
	Sad	3.3	16.7	15.4	30.4	5.6	20.8	7.7
	Fear	0.4	2.5	2.9	8.5	54.4	7.1	24.2
	Bored	3.8	17.3	12.5	22.1	6.5	29.2	8.8
	Surprise	1.7	14.2	4	7.5	26.3	8.1	38.3

		Anger	Neutral	Positive	Rest	Emphatic
FAU-Aibo	Anger	58.9	10.8	4.4	10	15.9
	Neutral	8.6	45.8	11.2	16.1	18.3
	Positive	3.2	21.8	54.8	17.6	2.7
	Rest	15.6	25.3	20.7	24.9	13.6
	Emphatic	15.2	19.5	3.8	9.5	51.9

The two-dimensional NDMS of our confusion matrices are shown in Figures (6-5, 6-6, and 6-7). Considering the NDMS graphs we adopt the Ens3 as our classification model, and design an MC classification for the three datasets as follows:

A. For Emo-Berlin dataset we combined the Anger and Happy emotions in a superclass SC={Anger, Happy}, therefore the first level classification contains the classes {SC, Neutral, Sad, Fear, Bored, Disgust}. While for the second level the classes involved in the classification model are {Anger, Happy}. Note that in this level the number of samples is less than 250 samples, and therefore choosing 250 PCs (as used in Ens3) is not applicable, and therefore we used 50,100,150 PCs in the MC instead of 50, 150, 250 PCs.

B. For the Kurdish dataset we defined CS1= {Surprise, Fear}, and SC2= {Bored, Sad}, therefore the regarded classes in the first level are {Anger, Happy, Neutral, SC1, SC2}. While in the second level a model is adopted to classify SC1 into Surprise and Fear, and another model to classify SC2 into Bored and Sad.

C. Finally for Aibo dataset the classes {Anger, SC, Positive, Emphatic} are classified in the first level, where SC is defined as {Rest, Neutral}, which will be clasiified in the second level.

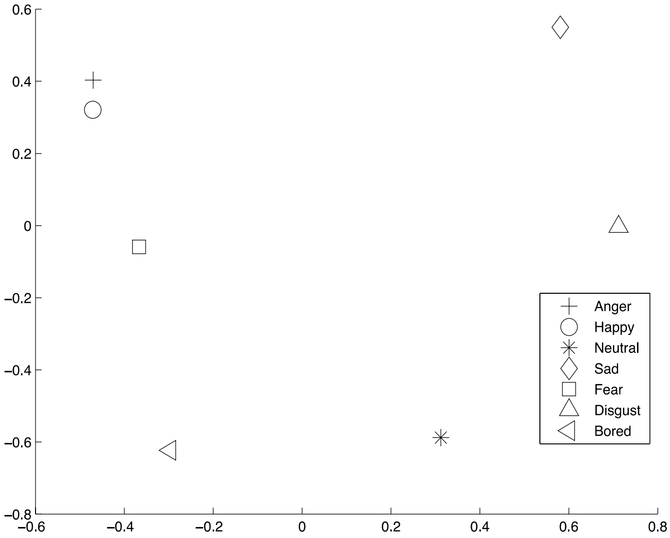

Figure 6-5: NMDS for Emo-Berlin dataset.

The recognition accuracy rate for the MC model is presented in table 6-7, and shows an improvement, albeit marginal, for both the Kurdish and Aibo datasets. The small number of samples in the Emo-Berlin dataset might be the reason for not improving the accuracy of MC over Ens3. To some extent in this case the MC is disadvantaged the small size of the samples available for training the 2nd level. However the MC model applied to Aibo dataset show significant improvement (with $p=0.048<0.05$) of the SER using Ens3. For the Kurdish dataset no significant improvement is observed at ($p=0.1>0.05$). This result in table 6-7 shows also that our suggested models outperform the state of art results for all the used datasets.

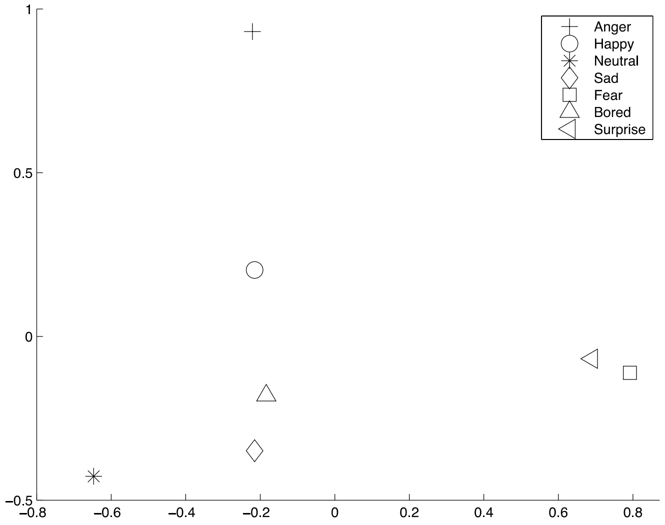

Figure 6-6: NMDS for the Kurdish dataset.

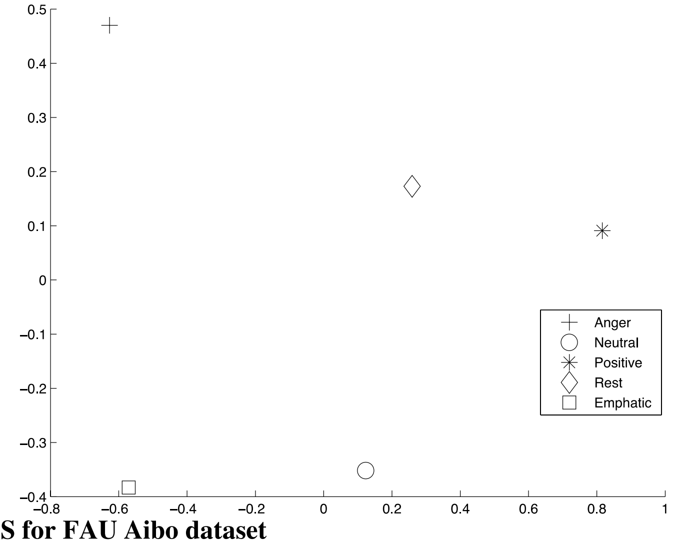

Figure 6-7: NMDS for FAU Aibo dataset

Table 6-6: Multi level Classifier & Ens3 accuracy rate comparison with state of the art results

	PCA+LDC	Ens3	MC	SOA
Emo-Berlin	90.7	**91.2**	89.5	89.9 (Vlasenko, et al., 2007)
Kurdish	43.5	44.9	**45.5**	**--**
Aibo	46.3	47.3	**47.9**	44 (Schuller, et al., 2011b)

The limited success of this pilot study is a motivation to conduct a more extensive investigation into MC SER schemes. This may require, a modified strategy in creating super classes.

6.5 Towards emotional content browsing in speech

The automatic score-based model evaluation of SER, done so far in this thesis, highlights the presence of a spectrum of various emotions in one speech portion, especially in the datasets that have high ratio of confusion between emotions. This has been consistently manifested in the various confusion matrices obtained throughout out the thesis, where we noted that only in the case of the Emo-Berlin dataset the confusion matrices had many zero entries, and there were no zeros in the confusion matrices for the other two datasets. The confusion matrix represents an aggregate of the accuracy using the scores from all the samples. Perhaps, a consideration of the scores for each sample may reveal more information on how the classification of each sample is reliable when the score of the sample for the other emotion is not much different to the accepted emotion class. The score-based decision reflects how far the score of the positive sample from the next negative one and judge the correct labelled samples, while ranking judges the incorrect labelled samples. The spider chart provides a simple visualization of the above reliability of decision concept, whereby the tested sample is placed at the center of a spider web while the emotions are places at the corners. The scores for a sample are marked along the central line in the direction of the emotion. Here we use the spider chart to visualize the scores for any tested sample toward the available emotions, using scores obtained from the LDC classifier. For example in testing one sample from the FAU-Aibo (see Figure 6-8 C), the implemented spider chart shows that the sample is more likely to belong to Anger emotion (55%), the second possible emotion is the Emphasized emotion (37%), next emotions are Neutral, Positive, and Rest emotions for just 4%, 2%, and 2% scores respectively. Similar interpretation is correct for spider charts in Figure 6-8 A&B.

The spider chart can be used to rank the sample with respect to the available emotions. Ranking based classification is a common practice in search engines, when searching for a specific term/object/item in text, music, speech, or by imaging. For instance Music Emotion Recognition (MER) is considerably studied in the research community because of the rich emotion content in music and human’s response to music listening is often related to emotion. Therefore, MER is considered as a promising means to enhance the organization, retrieval, and browsing of music collection (Yang & H. Chen , 2011). Similarly, emotion content browsing in speech seems to be promising filed in improving high quality search engines.

Ranking based emotion detection is helpful for emotion browsing, the aim being the detection and sorting the most similar instances to the browsed content. Emotional studies are rarely involved in applications of speech emotion ranking. However, emotion ranking was used in (Cao , et al., 2015) to improve the recognition accuracy of SER by adopting an emotion ranker using a pairwise SVM.

An appropriate tool for measuring the ability of ranking is the Receiver Operating Characteristic (ROC) curve (See Figure 6-9). ROC plots a curve of the true positive rate (sensitivity) against false positive rate (1- specificity) for a binary classification. The area under the ROC (AUC) measures the model performance of having true positive against the false positive decision. The AUC is given by the following formula:

$$A = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} 1_{x_i > y_j}}{mn}$$

Where $x_1, x_2, \dots , x_m$ are positive examples and $y_1, y_2, \dots, y_n$ are negative examples.

High AUC refers to the good ability of classes ranking, while bad recognition accuracy refers to choosing a bad cut-point threshold. The AUC present the quality of the model regardless of the cut-point used by the classifier, which allows the ignored scores above or below that threshold to be represented. Now AUC for the ROC curve is computed for PC+LDC scheme, described in (6.3) for one emotion versus all others.

In tables (6-8, 6-9, and 6-10) we present the confusion matrix, UAR, and AUC for each class against the rest of classes computed for the Emo-Berlin, FAU Aibo, and the Kurdish datasets. Significantly high AUC value is an indication of a good ranking quality of the model.

Although the PCA+LDC model is not a ranked-based classification but, generally the observed high accuracy is comparable to the AUC, unless the cut point/threshold of the classifier is chosen badly especially when the threshold is significantly biased away from the Equal Error Rates (EER). Ranking the emotion of the Emo-Berlin dataset is high for all emotions, and the worse AUC is observed for Anger emotion.

This is due to the ability of the participant in suppressing all emotions but the one that they are asked to act. For the FAU-Aibo dataset the rest emotion class seems to be poorly ranked and it is comparable to its recognition accuracy. Finally in the case of the Kurdish dataset Bored and Sad emotions have small AUC as an indication to the bad ranking of the model towards these two emotions.

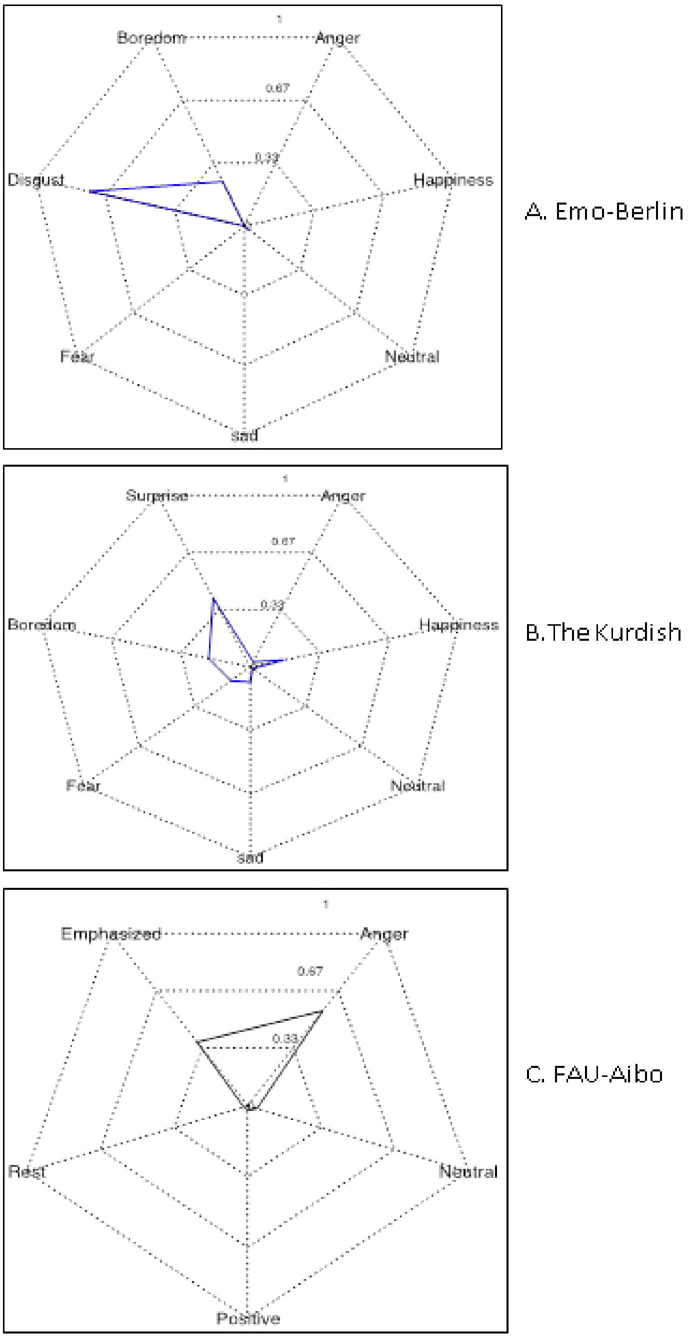

Figure 6-8: Spider Chart of one sample from each dataset

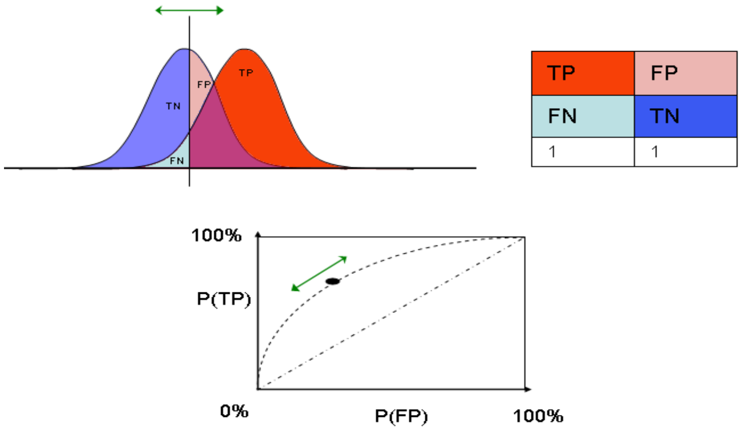

Figure 6-9: ROC relation to other measurements

Table 6-7: UAR, AUC for ROC curve for individual emotion vs. all, in Emo-Berlin database

EMO	Confusion Mat.		UAR	AUC
Angry vs. all	0.625	0.375	0.7412	0.8713
	0.14	0.8574		
Happy vs. all	0.9525	0.0474	0.8424	0.9289
	0.2676	0.7323		
Neutral vs. all	0.921	0.0789	0.9035	0.9656
	0.1139	0.886		
Sad vs. all	0.9852	0.0147	0.9926	0.9978
	0	1		
Fear vs. all	0.9785	0.0214	0.8370	0.969
	0.3043	0.6956		
Disgust vs. all	0.9779	0.0220	0.9704	0.9962
	0.0370	0.9629		
Bored vs. all	0.9897	0.0102	0.9079	98.81
	0.1739	0.8260		

Table 6-8: UAR, AUC for ROC curve for individual emotion vs. all, in FAU-Aibo database

Aibo	Confusion Mat.		UAR	AUC
Anger vs. all	0.7184	0.2815	0.7489	0.8175
	0.2206	0.7793		
Neutral vs. all	0.6729	0.3270	0.6663	0.7258
	0.3403	0.6596		
Positive vs. all	0.7631	0.2368	0.7281	0.7891
	0.3069	0.6930		
Rest vs. all	0.6281	0.3718	0.6126	0.6549
	0.40293	0.5970		
Emphatic vs. all	0.7195	0.2804	0.7096	0.781
	0.3003	0.6996		

Table 6-9: UAR, AUC for ROC curve for individual emotion vs. all, in Kurdish database

Kurdish	Confusion Mat.		UAR	AUC
Anger vs. all	0.7583	0.2416	0.8560	0.9475
	0.0461	0.9538		
Happy vs. all	0.7690	0.2309	0.6720	0.7381
	0.425	0.575		
Neutral vs. all	0.8329	0.1670	0.7727	0.8589
	0.2875	0.7125		
Sad vs. all	0.6902	0.3097	0.6263	0.6691
	0.4375	0.5625		
Fear vs. all	0.8517	0.1482	0.7706	0.8653
	0.3104	0.6895		
Bored vs. all	0.7024	0.2975	0.5887	0.6316
	0.525	0.475		
Surprise vs. all	0.7975	0.2024	0.6873	0.7848
	0.4229	0.5770		

6.6 Conclusion

To continue the work done in the last chapter, which was focused on dimension reduction techniques using SVM, the work in this chapter tries to overcome some drawbacks of classification models like SVM in a high dimensional emotional feature by adopting PCA+LDC. PCA+LDC is a simple scheme showing good performance for SER. High dimensional emotional attributes pre-processed by PCA produces informative de-correlated emotional meta-features, which show reasonable capability to facilitate the classification procedure for the simple classifier LDC. However, RP projections with the LDC classifier have not achieved what was achieved by the PCA+LDC, due to the wide dispersion of the data distribution in the random subspace.

The relatively modest number of experimentation with the ensemble of versions of PCA+LDC showed an improvement over the single PCA+LDC scheme. Those versions were based on training of selected subsets of samples or features. It was natural to investigate a new approach of the ensemble classifier using meta-feature subsets of the highest PCs (Ens3). This approach showed limited improvements over other ensemble schemes, and certainly over the state of art results, with (p=0.04, and p=1.9×10^{-10}, for both Emo-Berlin and Aibo dataset).

Consideration and visualization of the confusion matrices point towards the possibility of developing a multi-level classifier based on the entries in the confusion matrix, for improved SER accuracy. Indeed, the pilot study that used the Ens3 in a multi-level context we got improved recognition accuracy for the Aibo dataset. More research could reveal more insight into this problem/approach.

The spider charts is a visualization of a score based ranking of emotion classes, that could aid in gaining more knowledge to shed a light on the many SER influencing factors (e.g. speaker, culture, age, and gender). But this requires extensive investigations.

Chapter Seven
Conclusion

This thesis was devoted to investigate, develop and test the performance of automatic computational tools to deal with the main challenges in SER from a pattern recognition point of view. These investigations primarily targeted the main components of pattern recognition solutions starting with feature extraction, feature selection and dimension reduction, and ending with classification. Right at the outset, we found that SER belongs to a category of pattern recognition applications where there are wide disagreements among researchers or different interpretation of the main objects and concepts under investigation. Even the list of emotions is subject to debate, let alone how to determine the emotional state of speaker or determine factors that influence the recognition of emotions expressed by speakers. These basic difficulties are the main obstacle to determine the quantitative features/attributes that help develop a mathematical model for SER. Moreover, data collection and construction of speech based emotion-related database, necessary to test the performance of SER algorithms, is challenge in itself.

Our investigation into identifying speech signal features relevant to emotion recognition revealed a wide controversy in the literature on a limited set of features. Over the last two decades of SER research has led to the emergence of various types of acoustic features that can be extracted from the speech signal. In fact the non-convincing accuracy rate obtained by the suggested SER models especially for spontaneous datasets, encourage researchers to seek more and more features resulting in feature vectors of dimension that amount to several thousands. The obvious pattern recognition approach to deal with a high dimensional model is the use of deterministic/non-deterministic dimension reduction and meta-feature selection procedures, while most existing SER research tend to rely on simplistic, albeit credible, feature selection procedures. Beside this what so called "curse of dimension" problem is related to the lack of sufficient density of samples available for training.

In this thesis we dealt with the choice of adequate experimental datasets to test any recognition scheme to be developed, by conducting extensive experiments on three different datasets collected for different purposes. The first of these databases is an acted data by professional actors (Emo-Berlin) and cleaned from the unconvincing samples, the second dataset is acted by non-professional actors (the Kurdish dataset). The third one is the non-acted FAU-Aibo dataset that involved young children speakers (10-13 years). The use of various kinds of datasets is meant to provide evidences on the reliability and extendibility of achieved accuracy results to data that are collected under different conditions. Moreover, such experiments are expected to shed light on effect of variation in the ability of the participants in expressing specific emotion while speaking as well as variation in the reliability of the aftermath labelling of samples by a range of people of variable expertise.

The second issue that we confronted was the sufficiency, or otherwise, of currently researched sets of speech signal based emotion-related features. In chapter four we concluded that there are yet additional features, that are expected to have an impact on the expression of emotion during speech that may add information to the already high dimensional set of acoustic features. We proposed a set of features (ES) to be extracted from the LP-residual signal that models the excitation source of the speech signal, which is subjected to removing the influence of the vocal tract system. The ES is of relatively small dimension compared to the high dimensional features known as OP in this thesis. The experiments conducted in chapter 4 have demonstrated that the proposed ES features do indeed have complementary information of the OP features for improved accuracy.

The effect of various sets of features can be more highlighted by fusing them at the classification level. We evaluated weights for the decisions of models trained by ES and OP features using ANN and SVM classifiers. The fusion weights quantify the contribution of each feature sets to SER performance. For the developed weighted fused scheme we observed that ANN classifier is more effective by using it in a pairwise approach for SER. The pairwise network is simpler in terms of computation especially for high dimensional data, and is convenient for fusing with pairwise-based classification models like the SVM. However, the high dimensionality of the feature vectors (more than 6.5K dimension) raises reasonable questions about the possibility of the presence of redundant and correlated features/attributes. Although it is logical to assume the presence of redundancies, but proving this mathematically is not realistic due to the fact that we only know relatively small set of sample feature vectors whose attributes may be subject to computational errors. The alternative would be to use dimension reduction and feature selection to remove redundancies and/or to eliminate the influence of features that have minute/marginal contribution to accuracy. The ANN with one hidden layer (as used in chapter four), compress the input feature vector in the hidden layer, which is mapped later to one target output node. The ANN outperforms the SVM in the case of FAU-Aibo, and the Kurdish datasets, in which high confusion between classes is present, i.e. the accuracy rate for these two datasets is low (<50%). The high recognition accuracy obtained by ANN when using the high dimensional feature set (OP) encouraged and motivated the investigation of different dimension reduction methods in chapter 5.

We have investigated various approaches of dimension reduction in order to reduce the redundancy and correlation that is more likely to be present in high dimensional attributes obtained from a relatively small set of speech signal samples. Although, recent researchers suggested the usefulness of these approached for SER, the feature selection approach largely remained the most common way of dealing with curse of dimension. But adopted feature selection approaches are expected to leads to ignore vital information provided by the discarded features. Moreover, feature selection methods in a high dimensional feature space is either not efficient when using wrapper approach, or it might not be consistent with the classification technique when using filter approach. However, instead of dismissing feature selection approach altogether, we used feature selection methods that exploit the data-sparsity characteristics and a modified version of the SVM classifier, called DrSVM, which uses the l_1 minimization as an approximation for minimizing the cardinality of the set of the selected features. This approach was outperformed by, or had comparable performance to, the use of the whole set of features and even by the use of a random feature selection, which further encouraged seeking another alternative.

Based on the conclusions above, we tried to extract and select meta-features, which are attributes selected from the transformed space by different projection matrices like PCA and RP projections. The performances of these kinds of pre-processing techniques depend on the used classifier techniques. Our experiments have shown that data pre-processing by the Binary RP can have an adaptation role, which helps in reducing over-fitting when used with SVM classifier. But this is not the case when using the BRP matrices with the LDC classifier, due to the difficulty in estimating the statistical parameters in a random subspace, which are vital in building the LDC model. On the other hand, the PCA de-correlates the meta-features and thereby fitting the requirements of the LDC procedure that is based on the assumption that the classes' samples has a Gaussian distribution. Our investigations of dimension reduction techniques from more than 1K features using PCA or RP projections for SER, which has not been investigated so far; highlight the usefulness of meta-features in terms of efficiency and performance. Meta-features extracted by PCA and classified by LDC improve the results of the SER to 90.7%, 43.2%, and 46.3 for Emo-Berlin, the Kurdish and FAU-Aibo datasets respectively. Additionally PCA+LDC takes negligible time in model training.

Extracting and selecting meta-features using adaptive projections (like PCA) and non-adaptive ones (like RPs), was based on using the emotional dataset in training the PC projections or building projection matrices of random numbers which are totally independent from the data used in training the classifier. However, the question was what if the PC projection is trained using another emotional or Neutral speech samples. To answer such questions we conducted experiments to investigate the use of PC projections, trained by one emotional dataset or trained by datasets that are not relevant to emotions or speech, to classify data samples from another dataset. The PC projections trained by emotional speech samples that are not involved in classifying the data, followed by SVM classifier, show comparable accuracy rate to the PC projection of the data involved in the classification stage. And interestingly the PC projection matrices by Neutral speech samples also showed good capability in capturing the principal components of another emotional data. The Neutral class of emotion is assumed to be with score zero in the emotional primitives\dimensions like Arousal and Valence. Consequently, adequate number of Neutral samples shows usefulness in representing the correlations of emotional speech attributes. Although the use of PC projections trained on Mammogram image data did not achieve comparable accuracy, but their performance was far from disastrous. In conclusion, using PC's trained on non-speech data or even mix of different type of data require further investigations in the future.

The idea of extracting features from the emotional data is to present the more "representative" picture of the emotional speech samples. It would be useful to have different models that solve the same problem. To satisfy this target, we investigated the use of ensemble classifiers. The thesis consists of three forms of suggested ensembles that are based on selecting random subsets from the speech samples, random subsets from the feature sets, and finally using three subsets of meta-features extracted by PCA. The ensembles prove the possibility of improving the recognition accuracy rate, as an indication that each random subset from the features might draw a different picture of the emotional population, such that each of these representations are not linearly dependent i.e. they are non-redundant portions in the whole picture. There are many other possibilities of designing ensemble classifier like fusing various subsets of data and classifiers, which we expect to provide new avenues to be investigated in the future.

Despite the extensive experimentation and the various tested SER schemes a puzzling, though expected, observation continued to come to the fore. A huge gap in the SER accuracy rates are achieved by each of the emotion recognition schemes, including the SOA schemes, when tested on "acted" and on "non-prompted databases. We presented few arguments to explain the presence of this gap. But an important plausible explanation, supported by the results of our various investigations, is the presence of information relevant to more than one emotion in the same portion of a speech sample. This could be interpreted by the problem of labelling the emotional speech samples. For instance in a subjective test made for the Kurdish dataset the correct labelling of the emotional recordings by 10 listeners was just 41%. Any pattern recognition model is expected to achieve accuracy rate close to what obtained by human capability. We can conclude that the speech signal is not sufficiently adequate source of information about the full picture of emotions. In this respect we note that the human ability of recognizing emotions from the speech signal uttered by others is reported frequently to be non-convincing.

The various experimental work, conducted in this thesis, together with experience gained by researchers in data mining and pattern recognition provide convincingly credible argument that this performance gap in the accuracy of emotion recognition from speech could not be dealt with as a simple classification problem. In other words speech is not sufficient for emotion recognition, and the same speech portion maybe modulated by a spectrum of different emotions, especially in the non-prompted datasets. Our argument is based on the fact that these databases are recorded under different circumstances and for different purposes. In general, determining the emotion of a person from a recorded speech, by human observers, is a very difficult task without access to facial expressions and/or bodily gestures during the uttering of the speech. This is exactly what happened when recording the non-prompted databases, which were supposed to capture emotions from the real life speech. This difficulty explains the rather low accuracy rates for the FAU Aibo and the Kurdish datasets. Although the Kurdish database is acted database; but the speaker are told to act the emotions using their own style, and all of the produced sentences have been involved in the study without removing the unconvinced ones. In the case of the acted Emo-Berlin datasets the subjects obviously attempt to suppress all but one emotion that they are asked to stress, and consequently all the schemes achieve significantly higher accuracy. A sceptic may refute this last argument by pointing out that the best accuracy rate is a mere 91.2%. However, we argue that achieving higher accuracy over prompted/acted datasets is certainly possible but this requires the speakers to be professional actors of higher standard than participants in existing databases. In fact, the Emo-Berlin database was designed for emotion synthesis purpose (not recognition), and according to the documentation of the database about 250 samples out of original 800 recording samples have been removed from the dataset due to variation in expert listeners judgment.

The above discussion, together with the observations depicted in the presenting the model recognition scores in terms of Spider charts, motivate our hypothesis, to be investigated in the future, that emotion recognition from speech should not be dealt with as a simple classification problem. In other words speech is not sufficient for emotion recognition, but the same speech portion maybe modulated by a spectrum of different emotions, especially in the non-prompted scenarios.

Along the investigation made in this thesis, we recommend to further investigate SER with regards to the following issues in the future to shed more light on the various challenges discussed in this thesis:

1. The use of boosting based ensemble classification; thereby incorporate various classifier techniques in building an SER. To avoid possible over-fitting that may occur by the boosting approach in the context of high dimensional feature vectors, we recommend the use of a hybrid ensemble classifier that include the bagging approach by having various random

subsets of features each subset is classified by a number of classifiers in a boosting model.

2. The investigation made in chapter 6, refer to the possibility of using Speech Emotion ranking for the promising application of Speech Emotion Browsing (SEB). This approach is also requiring further investigations.
3. During the extensive speaker independent based SER experiments in this thesis, we have observed different level of contributions to SER accuracy for different groups of speakers. This observation encourages investigating the "Doddington Zoo" problem, which is well-known problem in biometric area. Here, the speakers would be marked as sheep, goat, lamb, or wolf based on the speaker samples contribution in the SER accuracy in terms of their confusion matrix. We recommend this investigation to have more information about the speaker "ability" of expressing emotion. This could also lead to developing speaker-dependent SER schemes.
4. For more reliable emotion recognition system, it is necessary to focus on multimodal approach to be fed from various sources of information like face, speech and gesture.

Bibliography

Achlioptas, D. (2003), *Database-friendly random projections: Johnson-Lindenstrauss with binary coins*, Computer and System Sciences , Volume 66, pp. 671–687.

Anill, K. & Robet, P. W. (2000). *Statistical Pattern recognition: A review,* IEEE transaction on Pattern analysis and machine intellegence, 22(1), pp. 4-37.

Bach, F., Jenatton, R., Mairal, . J. & Obozinski, G. (2011), *Optimization with Sparsity-Inducing Penalties*, Foundations and Trends in Machine Learning, 4(1), pp. 1–106.

Bai, J. (2011). *Estimating High Dimensional Covariance Matrices and its Applications*, Annals of Economics And Finance, 12(2), pp. 199-215.

Batliner, A., Steidl, S. & Noth, . E. (2008a), *Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus,* Marrakesh, Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect. pp. 28-31.

Batliner, A., Schuller, B., Schaeffler, S. & Steidl, S. (2008b), *Mothers, adults, children pets: towards the acoustics of intimacy,* Las Vegas, IEEE- ICASSP, p. 4497 - 4500.

Batliner, A., Stiedl, S. & Noth, E. (2008c). *Private Emotion vs. Social intraction a Data driven Approach toward Analysis Emotion in Speech*, User Modelling and User-Adapted Intraction (umani), 18(1-2), pp. 175-206.

Batliner , A. et al. (2011), *Whodunnit – Searching for the most important feature types signalling emotion-related user states in speech*, Computer Speech & Language, 25(1), pp. 4 - 28.

Bermejo, P., de la Ossa, L., Gamez, J. & Puerta, J. (2012). *Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking,* Knowledge-Based Systems, pp. 35-44.

Beyer , K. S., Goldstein, J., Ramakrishnan, R. & Shaft, U.(1999), *When Is "Nearest Neighbor" Meaningful?,* London, UK, 7th International Conference on Database Theory, pp. 217-235.

Bingham, E. & Mannila, H. (2001), *Random projection in dimensionality reduction: applications to image and text dat,* New York, NY, USA, 7th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 245-250.

Buckow, J. et al. (1999), *Fast and robust features for prosodic classification*, In: V. Matousek, P. Mautner, J. Ocelíková & P. Sojka, eds. *Text, Speech and Dialogue.* pp. 193-198.

Burkhardt, F. et al. (2005). *A database of german emotional speech.* Lissabon, Interspeech, pp. 1517-1520.

Candes, E. J. & Tao, T. (2005), *Decoding by linear programming*, Information Theory, IEEE Transactions on, 15(12), pp. 4203-4215.

Candès , E. J. & Wakin, M. B. (2008), *An Introduction To Compressive Sampling,* IEEE Signal Processing Magazine, MARCH , 25(2), pp. 21-30.

Cao , H., Verma , R. & Nenkova, A. (2015), *Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech*, Computer Speech & Language, 29(1), pp. 186 - 202.

Carlos , R., Vladik , K. & Miguel , A. (2013), *Why ℓ1 Is a Good Approximation to ℓ0: A Geometric Explanation*, Journal of Uncertain Systems, 7(3), pp. 203-207.

Chauhan, A., Koolagudi, S. G. & Kafle, S. (2010), *Emotion Recognition using LP Residual,* IIT Kharagput, IEEE Students' Technology Symposium, pp. 255-261.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2012), *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, Volume 16, pp. 321–357.

Chen , L., Mao , X., Xue , Y. & Lung , L. (2012), *Speech emotion recognition: Features and classification models*, Digital Signal Processing, 22(6), pp. 1154–1160.

Cortes, C. & Vapnik, V. (1995), *Support-Vector Networks*, Machine Learning, 20(3), pp. 273-297.

Cover, T. & Hart, P. (1967), *Nearest neighbor pattern classification*, Information Theory, IEEE Transactions, 13(1), pp. 21-27.

Cowie, R. et al. (2001a), *Emotion recognition in human–computer interaction*, IEEE Signal Process. Mag., 18(1), pp. 32–80.

Cowie, R., Sussman, N. & Ben-Ze'ev, A. (2001b), *Emotion: Concepts and Definitions*, In: Emotion-Oriented Systems: The HUMAINE Handbook, Verlag Berlin(Heidelberg): Springer, pp. 9-30.

Dalal , N. & Triggs , B. (2005), *Histograms of oriented gradients for human detection*, IEEE-Computer Vision and Pattern Recognition, pp. 886-893.

Daniel , N., Kjell , E. & Kornel , L. (2006), *Emotion Recognition in Spontaneous Speech Using GMMs,* Pittsburge, PA, USA, Interspeech, pp. 809-812.

Dasgupta, S. (1999), *Learning Mixtures of Gaussians*, IEEE Symposium on Foundations of Computer Science, PP. 634-644.

Dietterich, T. G. (1997), *Machine Learning Research: Four Current Directions*, Artificial Intelligence Magzine, Volume 4, pp. 97-105.

Dietterich, T. G.(2000), *Ensemble methods in machine learning*, In: Multiple classifier systems, Springer Berlin Heidelberg, pp. 1-15.

Douglas-Cowie, E. et al. (2005), *Multimodal Databases of Everyday Emotion: Facing up to Complexity,* Lisbor, Portugal, Interspeech, pp. 813-816.

El Ayadi, M., Kamel , M. S. & Karray, F. (2011). *Survey on speech emotion recognition: Features classification schemes and databases*, Pattern Recognition, 44(3), pp. 572 - 587.

Elfenbein, H. A. & Ambady, N. (1986), *On the universality and cultural specificity of emotion recognition: A meta-analysis*, Psychological Bulletin , 28(2), pp. 203–235.

Emmanuel, J. & Wakin, M. B. (2008), *An introduction to compressive sampling*, Signal Processing Magazine, *IEEE,* 25(2), pp. 21-30.

Engberg, I. S. & Hansen, A. V. (2007), *Documentation of the Danish Emotional Speech Database,* Internal AAU report, Center for Person Kommunikation, Denmark, 22.

Eyben, F. et al. (2010a), *Cross- corpus classification of realistic emotions: some pilot experiments*, Valletta, Malta, 7th International Conference on Language Resources and Evaluation (LREC), pp. 77–82.

Eyben, F., Wöllmer, M., & Schuller, B. (2009), *OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit*, In Affective Computing and Intelligent Interaction and Workshops, 2009, ACII 2009, 3rd International Conference on, pp. 1-6.

Eyben, F., Wöllmer, . M. & Schuller, B. (2010b), *openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor,* Firenze, Italy, ACM Multimedia (MM), ACM, pp. 25-29.

Fant, G. (1970), *Acoustic theory of speech production,* Mouton, The Hague, Paris, Volume 2.

Farrús, M., Hernando, J. & Ejarque, P. (2007), *Jitter and Shimmer Measurements for Speaker Recognition.* Proceeding of Interspeech, pp. 778-781.

Fisher, R. A. (1936), *The use of multiple measurements in taxonomic problems*, Annals of eugenics*,* 7(2), pp. 179-188.

Gobl , C. & Chasaide, A. N. (2003), *The role of voice quality in communicating emotion, mood and attitude,* Speech Communication, 40(1-2), pp. 189 - 212.

Hansen , J. & Bou-Ghazale , S. (1997), *Getting started with susas: A speech under simulated and actual stress database*. Rhodes, Greece, ISCA-Eurospeech, pp. 1743–1746.

Hansen , J. H. & Cairns, D. A. (1995), *ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments*, Speech Communication*,* 16(4), pp. 391 - 422.

Hassan, A., & Damper, R. (2010), *Multi-class and hierarchical SVMs for emotion recognition*, Proceedong of Interspeech 11th, pp. 2354-2357.

Hassan, A. & Damper, R. (2012), *Classification of emotional speech using 3DEC hierarchical classifier*. Speech Communication, Volume 54, pp. 903-916.

Hassan , A. & Damper, R. (2013), *On Acoustic Emotion Recognition: Compensating for Covariate Shift*, IEEE Transaction on Audio, Speech, and Languege Processing, July, 21(7), pp. 1458-1468.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning,* 2nd ed. New York: Springer New York Inc.

Heaton, J. (2008), I*ntroduction to Neural Networks for Java*, Heaton Research.

Hecht-Nielsen, R. (1994), *Context vectors: general purpose approximate meaning representations self-organized from raw data*, Computational Intelligence: Imitating Life, pp. 43–56.

Hsu, C.-W., Chang, C.-C. & Lin, C.-J. (2010), *A Practical Guide to Support Vector Classification,* Tech. rep., Department of Computer Science, National Taiwan University.

Johnson , W. B. & Lindenstrauss, J. (1984), *Lipshitz mapping into Hilbert space,* Modern Analysis and Probability, pp. 189-206.

Jolliffe, I. T. (2002), *Principal Component Analysis*, Second ed. New York: Springer.

Kanade, T., Cohn, J. F. & YingLi , T. (2000), *Comprehensive database for facial expression analysis*, 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46-53.

Kojadinovic, I. & Wottka, T. (2000), *Comparison between a filter and a wrapper approach to variable subset selection in regression problems,* European Symposium on Intelligent Techniques (ESIT).

Koolagudi, S. G. & Rao, K. S. (2012a), *Emotion recognition from speech using source, system, and Prosodic features*, Intrenational Speech Technology, 15(2), pp. 265-289.

Koolagudi, S. G. & Rao, K. S. (2012b), *Emotion recognition from speech: a review*, International Journal of Speech Technology, 15(2), pp. 99–117.

Lee , C. et al. (2011), *Emotion recognition using a hierarchical binary decision tree approach*, Speech Communication*,* 53(9), pp. 1162–1171.

Lee, C. M. & Narayanan, S. S. (2005), *Towards detecting emotions in spoken dialogs*, *IEEE Transactions on Speech and Audio Processing,* 13(2), pp. 293-303.

Leinfelder, H (1979), *A geometric proof of the spectral theorem for unbounded self-adjoint operators*, Mathematische Annalen, 242(1), pp. 85-96.

Liu, D., Qian, . H., Dai, G. & Zhang, Z. (2013), *An iterative SVM approach to feature selection and classification in high-dimensional datasets*, Pattern Recognition*,* 46(9), pp. 2531–2537.

Long Pao , T., Chen, Y.-T., Heng Yeh , J. & Hao Chang , Y. (2005), *Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification,* Tainan, Taiwan, The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), pp. 203-212.

Malandrakis , N., Potamianos , A., Evangelopoulos , G. & Zlatintsi , A. (2011), *A Supervised Approach To Movie Emotion Tracking*, Prague, IEEE-Acoustics, Speech and Signal Processing (ICASSP), pp. 2376 - 2379.

Martin, O., Kotsia, I., Macq, B. & Pitas , I. (2006), *The eNTERFACE'05 Audio-Visual Emotion Database,* Data Engineering Workshops, Proceedings, 22nd International Conference on IEEE, pp. 8-8.

Mika, S. et al. (1999), *Kernel PCA and De-Noising in Feature Spaces*, MIT Press-Advances In Neural Information Processing System, pp. 536-542.

Moon, H. & Phillip, J. P. (2001), *Computational and performance aspects of PCA-based face-recognition algorithms*, Perception-London, 30(3), pp. 303-322.

Murray, I. & Arnott, J. (1993), *Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion*, The Journal of the Acoustical Society of America, 93(2), pp. 1097–1108.

Muto, M., Kato, H. & Tsuzak, M. (2005), *Effect of speaking rate on the acceptability of change in segment duration*, Speech Communication, 74(3), pp. 277 - 289.

Ortony, A. & Turner, T. J. (1990), *What's Basic About Basic Emotions?*, Psychological Review, 97(3), pp. 315-331.

Oster, A. & Risberg, A. (1986), *The identification of the mood of a speaker by hearing impaired listeners*, STL-QPSR, 27(4), pp. 79-90.

Ou, G. & Murphey, Y. L. (2007), *Multi-class pattern classification using neural networks*, Pattern Recognition, 40(1), pp. 4-18.

Paliwal, K. K. (1998), *Spectral subband centroid features for speech recognition,* IEEE- Acoustics, Speech and Signal Processing, pp. 617-620.

Pao , T., Chen , Y., Yeh , J. & Chang , Y. (2005), *Emotion recognition and evaluation of Mandarin speech using weighted D-KNN classification,* Conference on Computational Linguistics and Speech Processing XVII, pp. 96-105.

Paulo, P. M., Larry, E. B. & Leslie, S. G. (1999), *Emotion Recognition in Psychotherapy: Impact of Therapist Level of Experience and Emotional Awareness*, Journal Of Clinical Psychology, 55(1), pp. 39-57.

Pérez, H., Carlos, R. A. & Luis, V. (2012), *Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model*, Biomedical Signal Processing and Control, 7(1), pp. 79-87.

Petrushin, V. A. (1999), *Emotion in Speech: Recognition and Application to Call Centers*, Artificial Neural Networks in Engineering, pp. 7-10.

Picard , R. W., Vyzas , E. & Healey , J. (2001), *Toward Machine Emotional Intelligence: Analysis of Affective Physiological State*, IEEE Transactions On Pattern Analysis And Maghine Intellegence , 23(10), pp. 1175-1191.

Pudil, P., Novovičová, J. & Kittler, J. (1994), *Floating search methods in feature selection*, Pattern Recognition Letters, Nov., 15(11), pp. 1119-1125.

Rabiner , L. & Juang , B. (1993). *Fundamentals of Speech Recognition.* 1st ed.

Rabiner, L. R. & Schafer, R. W. (2007), *Introduction to Digital Speech Processing*, Foundations and Trends in Signal Processing, 1(1-2), p. 1–194.

Rahimi , A. & Recht , B.(2007), *Random Features for Large-Scale Kernel Machines,* In Advances in neural information processing systems, pp. 1177-1184.

Scherer, K. R. (2000), *Psychology models of emotion*, In: J. Borod, ed. The neuropsychology of emotion. Oxford University press ed. pp. 137-166.

Scherer, K. R., Banse, R. & Wallbott, H. G. (2001), *Emotion Inferences From Vocal Expression Correlate Across Languges and Cultures*, Journal of Cross-Cultural Psychology, 32(1), p. 76–92.

Schipor , O. A., Pentiuc , S. G. & Schipor , M. D. (2011), *Towards a Multimodal Emotion Recognition Framework to Be Integrated in a Computer Based Speech Therapy System*, Brasov, Speech Technology and Human-Computer Dialogue, pp.1-6.

Schuller, B., Müller, R., Lang, M. & Rigoll, G. (2005a), *Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles.* Interspeech, ISCA, pp. 805-809.

Schuller, B. et al. (2005b), *Speaker independent speech emotion recognition by ensemble classification*, IEEE International Conference on Multimedia and Expo, ICME, pp. 864-867.

Schuller, B. et al. (2007), *The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals,* Antwerp, Belgium, interspeech, pp. 2253-2256.

Schuller, B. & Batliner, A. (2009a), *The Interspeech 2009 Challenge*, Brighton UK, Intrespeech 09, pp. 312-315.

Schuller, B., Vlasenko, B. & Florian , E. (2009b), *Acoustic Emotion Recognition: A Benchmark Comparission of Performance*, IEEE Workshop on Automatic Speech Recognition & Understanding.

Schuller, B. et al. (2011a), *Interspeech 2011 speaker state challenge,* Florance, Italy, 12th Annual Conference of the International Speech Communication Association, pp. 3201-3204.

Schuller, B., Batliner, A., Steidl, S. & Seppi, D. (2011b), *Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge*, Speech Communication, 53(9), p. 1062–1087.

Shuhong, J., Gesen, Z., Ping, C. & Xiaoli, X. (2010), *Bernoulli Compressed Sensing and Its Application to Video-Based Augmented Reality*, Journal of Computational Information Systems, 6(14), pp. 4819-4826.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Volume 26, CRC Press.

Simmons, S., Peng, . J., Bienkowska, J. & Berger, B. (2015), *Discovering What Dimensionality Reduction Really Tells Us About RNA-Seq Data*, Journal of Computational Biology, 22(8), pp. 1-14.

Slaney, M. & McRoberts, G., (1998), *Baby Ears: A Recognition System For Affective Vocalizations*, IEEE-International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1-4.

Solomon , R. (1980), *The opponent-process theory of acquired motivation: the costs of pleasure and the benefits of pain*. American psychologist, 35(8), pp. 691.

Soma, K., Joyanta, B. & Milton S., B. (2012), *Performance evaluation of pbdp based real time Identification system with normal MFCC vs MFCC of residual features*, Perception and Machine Intelligence, Lecture Notes in Computer Science, pp. 358-366.

Steidl, S. (2009), *Automatic classification of emotion-related user statesin spontaneous childern's speech*, PhD thesis, Depatrment of Computer Science, University of Erlangen-Nuremberg, Germany.

Vlasenko, B., Schuller, B. & Wen, A. (2007), *Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech*, Antwerp, Belgiume, Interspeech.

Vogt , T., Wagner, J. & André, E. (2008), *Automatic Recognition of Emotion from Speech: A Review of the Literture and Recommendtions for Practical Realisation*, Affect and emotion in HCI, pp. 75-91.

Vogt, T. & Andre, E. (2005), *Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognitio,*. IEEE International Conference on Multimedia and Expo-ICME, pp. 474 - 477.

Wang , L., Zhu , J. & Zou, H. (2006), *The doubly regularized support vector machine*, Statistica Sinica, 13(2), pp. 589-616.

Webb, A. R. (2002), *Statistical Pattern Recognition,* Second ed. Malvern, John Wiley & Sons.

Williams, C. E. & Stevens, K. N. (1972), *Emotions and Speech: Some Acoustical Correlates*, The Journal of the Acoustical Society of America, 52(4), pp. 1238-1250.

Yang , Y.-H. & H. Chen , H. (2011), *Ranking-Based Emotion Recognition for Music Organization and Retrieval*, IEEE Transactions On Audio, Speech, And Language Processing, 19(4), pp. 762-774.

Yang, B. & Lugger, M. (2010), *Emotion recognition from speech signals using new harmony features*, Signal Processing, 90(5), p. 1415–1423.

Yessad , D. & Amrouche , A. (2012), *SVM based GMM Supervector Speaker Recognition using LP Residual Signal*, Lectures Notes in Electrical Engineering, June.pp. 579-586.

Yeung, K. Y. & Ruzzo, W. L. (2001), *Principal component analysis for clustering gene expression data*, Bioinformatics, 17(9), pp. 763-774.

Yoo, S. H., Matsumoto, D. & LeRoux, J. A. (2006), *The influence of emotion recognition and emotion regulation on intercultural adjustment*, International Journal of Intercultural Relations, 30(3), p. 345–363.

Yu, L. & Liu, H. (2003), *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution,* 20th International conferance of Machine Learning, pp. 856-863.

Zhang , S. & Zhao , X. (2013), *Dimensionality reduction-based spoken emotion recognition*, Multimedia Tools and Applications, April, 63(3), pp. 615-646.

Zhang, Z. & Schuller , B. (2014), *Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition*, IEEE Signal Processing Letter, 21(19), pp. 1068-1072.

Zheng, N., Lee, T. & Ching, P. C. (2007), *Integration of Complementary Acoustic Features for Speaker Recognition*, Signal Processing Letters, IEEE*,* 14(3), pp. 181-184.