UNIVERSITY OF LEICESTER

Thesis submitted for the degree of

Doctor of Philosophy in Computer Science

Automatic Analysis of Voice Emotions in Think Aloud Usability Evaluation: A Case of Online Shopping

by

Samaneh Soleimani

Department of Informatics

February 8, 2019

Abstract

Emotions elicited from interacting with technologies are fundamental components of user experience (UX). Two general approaches to measuring people's emotional experiences exist: memory-based and moment-based. Whereas memory-based approaches are susceptible to the peak-end effect, moment-based approaches appear to better reflect the actual experiences of emotions. In this thesis, it is proposed that assessment of emotions of think aloud verbalisations is a moment-based approach for measuring emotional experiences. To evaluate the effectiveness of this approach, two independent user studies were conducted, respectively with 46 and 35 participants, in the domain of online shopping. In both studies, three assessment methods were applied for measuring emotional experiences: Self-reports, Vocal expressions (manual analysis), and Vocal Expressions (automatic analysis).

Study 1 adopted machine learning (ML) approaches to analyse participants' vocal expressions elicited during thinking aloud based on the discrete and dimensional models of emotions. While vocal analysis of verbal data was used as the moment-based approach, self-report questionnaires served as a means for the memory-based assessment. Results suggested that retrospective evaluations of emotions were significantly correlated with the most frequently elicited emotion (modal emotion) during the interaction.

Study 2 expanded the extent of analysis of vocal expressions by taking a different set of ML techniques. Emotions were modelled by dimensional descriptors through binary pleasure (negative/positive), binary arousal (low/high) and ternary dominance (low/neutral/high). Results showed 66%, 70%, 41% recognition accuracy for Pleasure, Arousal and Dominance dimensions respectively. Moreover, elicited facial expressions were analysed to derive classification models to predict subjective self-reports.

The main contribution of this thesis is the proposition and validation of an approach for automatic assessment of emotional experiences evoked during the think aloud protocol. Future work will investigate methods for improving the accuracy of automatic analysis, integration of multiple modalities (verbal and nonverbal) and different ML techniques in an actual automatic assessment tool.

Acknowledgements

I would like to start by expressing my sincere gratitude to my supervisor Professor Effie Law. Her continuous support, guidance and feedback inspired and shaped this research. She fostered curiosity and enthusiasm in me and taught me how to view science through research. This PhD would have not been possible without her constructive supervision. In addition, I am particularly thankful for her help in sourcing funding for my PhD and thus allowing me to pursue my academic dream here in the UK.

I would also like to thank Dr Leandro L. Minku, my second supervisor, for his great support and input on my research. I am sincerely grateful to Dr Fer-Jan de Vries, the postgraduate tutor at the Informatics department of the University of Leicester, who listened to me compassionately when I needed help. I will always think of him as a great person. My thanks also to Professor Thomas Erlebach for his enormous support throughout my stay in Leicester, to Dr Paul Rudman for his feedback on my work and to Professor Rick Thomas for his kind support. I am eternally grateful to everyone who took the time out of their busy schedules and participated in my empirical studies; your input made my thesis possible.

Moreover, I would like to thank all my friends and fellow PhD students who created beautiful experiences for me in Leicester. Your help and support motivated me to the very end of this work. Amongst everyone, I would like to particularly thank Matthias, my true friend and colleague, Gabriela and Stephanie, who greatly influenced my stay in Leicester.

Finally, I would like to thank my amazing family for being there to help me pursue my dreams. My mother, my role model, who patiently supported me through this long journey, filling me with the determination needed to never lose hope and give up. My father and sisters, Sareh and Saeedeh, who supported and looked after me every single day. And my deepest appreciation to my friend and hero, Omid, who stood by me from the very beginning to the very end. His constant patience, encouragement and love supported me in bringing this work to completion.

Dedications

I dedicate this thesis to my parents, my amor, my sisters and to whom supported me to be where I am.

Contents

List of Figures

List of Tables

Abbreviations

AAAC Association for the Advancement of Affective Computing.

AU Action Unit.

CNN Convolutional Neural Network.

DES Differential Emotion Scale.

DL Deep Learning.

DV Dependent Variable.

EC Emotion Challenge.

EDA Electrodermal Activity.

EEG Electroencephalography.

EMG Electromyography.

ESM Experience Sampling Method.

FACS Facial Action Coding System.

FN False Negative.

FP False Positive.

G-Mean Geometric Mean.

GSR Galvanic Skin Response.

GUI Graphical User Interface.

HCI Human-Computer Interaction.

HR Heart Rate.

IV Independent Variable.

KNN K-Nearest Neighbour.

LLD Low-Level Descriptors.

ML Machine Learning.

PAD Pleasure-Arousal-Dominance.

RGB Red, Green, Blue.

RNN Recurrent Neural Network.

RQ Research Question.

S-O-R Stimulus-Organism-Response.

SAM Self-Assessment Manikin.

SD Standard Deviation.

SDK Software Development Kit.

SMO Sequential Minimal Optimization.

SNS Somatic Nervous System.

STM Short-Term Memory.

STR Speech-to-Text-Recognition.

SVM Support Vector Machine.

TN True Negative.

TP True Positive.

UAR Unweighted Average Recall.

UX User Experience.

WAR Weighted Average Recall.

Glossary

Accuracy The degree to which the result of a measurement (such as recognition) conforms to the correct value.

Acoustic feature Vocal parameters or elements present in a voice/speech sound and capable of being experimentally observed, recorded, and reproduced.

Emotional Corpora A collection of material (such as audio, text, video, etc.) in machine-readable format assembled for the purpose of emotion analysis research.

Model A model is a classification algorithm that is trained based on the training dataset using a supervised learning method.

Precision Is the ratio of relevant instances among the recognised instances (also known as specificity).

Recall Is the ratio of relevant instances that have been recognised over the total amount of relevant instances (also known as sensitivity).

Segment Any discrete unit of a sequence meaningful to the field of analysis.

Speech What is spoken or uttered as in written words.

System System is defined as any type of interactive digital artefacts (products or services) such as applications, software, or web pages.

Testing Data The data that is independent of the training data and machine learning models are tested on.

Training Data The data with known labels that is used for training machine learning classifiers.

Utterance A semantically and linguistically well-defined unit of speech.

Verbalisation The process of gathering spoken data.

Voice The sound produced by the vocal organs of a vertebrate, especially a human.

Chapter 1

Introduction

1.1 Motivation

The root of user experience (UX) in psychological concepts such as happiness, frustration, and other emotions has intrigued the Human-Computer Interaction (HCI) research community to enquire the measurability of UX (Law et al., 2014). The controversy on the measurability of UX is derived from the ongoing debates on whether the experiential qualities of UX, such as pleasure, surprise and beauty could be broken down into a number of measurable metrics or not. In other words, the major argument has been whether UX experiential qualities should be studied through quantitative or qualitative approaches.
Adopting Hand's definition of measurement (Hand, 2004), where quantification has been defined as "the assignment of numbers to represent the magnitude of attributes of a system we are studying or which we wish to describe" (p. 3), experiential qualities should be described in a valid, reliable, useful and meaningful way. Thus, the significant value of quantitative measurements cannot be overlooked when it comes to UX measurement. On the other hand, qualitative methods have the advantage of providing useful insights into the causes of certain experiences. As a consequence, a combination of qualitative and quantitative approaches can constitute a true picture of UX, and thus it has been the preferable approach for many researchers in measuring UX (Bargas-Avila and Hornbæk, 2011; Law et al., 2009).

The prevalent UX evaluation methods have been borrowed mainly either from usability (such as questionnaires) or from psycho-physiological techniques (such as sensor-based approaches). This is due to the fact that UX has developed from usability, but encompassing complex concepts such as emotions. However, both usability and psycho-physiological techniques have some limitations when employed in UX studies (which will be elaborated in Section 2.2). In addition, considering the importance of both qualitative and quantitative approaches, the challenge is to develop new instruments and techniques that can inherit the strengths of the traditional measurement methods as well as tackle their limitations. In other words, the development of new approaches for UX does not imply abandoning the traditional usability techniques. Conversely, the usability approaches can be served as the ground for new UX measurement methods to be developed upon. The significance of developing reliable and easy to use UX measurement methods is especially high for industry, where systematic evaluation is part of the entire design process (Wixon, 2011).

As the notion of UX is defined as subjective, context-dependent and dynamic (Law et al., 2009), a UX measurement method should be self-reported, adaptive and trajectory-based (Law and van Schaik, 2010). Usability methods such as questionnaire, interview, and think aloud approaches can be used for capturing self-reported data. Self-reported data supplemented by qualitative content have the advantage of adding information about reasons behind actions and preferences; and thus can derive insight into the context and dynamism of UX. Nonetheless, users' emotions, the integral concept of UX, are short-lived (Scherer, 2005) and are difficult to be reproduced exactly to the same reactions they were elicited initially (Hazlett, 2006). The short duration of emotions highlights the necessity of employing trajectory and moment-based approaches. Psycho-physiological approaches allow the capture of the dynamic state of how bodily responses change over time, although without an explicit link to the context of elicitation (Wixon, 2011). Due to the fact that emotions are appraisal-driven or in other words triggered as responses to the cognitive evaluation of events (Scherer et al., 2001), hence understanding the cause of their stimulation is essential. This is specifically pertinent to interactive experiences, where decisions on whether and how to change a product/service to be made by designers and developers, who should take users's emotional reactions into account.

Emotion has been recognised as one of the main components in many cognitive processes in human-human or human-machine interactions such as learning and decision making (Goleman, 2006; Schwarz, 2000). Affective computing, pioneered by Rosalind Picard (Picard and Picard, 1997), denotes a research area focusing on the development of computers that are enabled to recognise emotions from observable signals and to respond to them accordingly. Since the advent of this notion, automated analysis of human emotions has been increasingly investigated by researchers from different disciplines such as computer science, linguistics and engineering. In affective computing, sensory data is recorded, processed and classified into discrete classes (such as emotions) using supervised Machine Learning (ML) algorithms.

Consequently, interpreting users' emotions is the common objective for both UX and affective computing research areas. Therefore, advancements in recognition of emotions could also serve the usability and UX community as well, where understanding emotional trajectories helps designing systems that can create positive experiences, which is the main focus of UX.

In this research I propose and investigate a methodological approach to assess emotional experiences during the course of interaction with technology. This methodology constitutes 1) the utilisation of the think aloud protocol (Section 2.2.1) - the widely used approach in usability testing and 2) the combination of the ML techniques (Chapter 3). Specifically this research investigated the simultaneous employment of the think aloud protocol and recognition of emotions expressed nonverbally and through vocal cues in a study presented in Chapter 4, and the combination of vocal and facial expressions in a second study (Chapter 5) for measuring emotional experiences elicited as results of interaction with technology.

In this methodology, it is assumed that the continuous assessment of emotions expressed during thinking aloud verbalisations is congruent with the moment-based approach of Kahneman in measuring instantaneous experiences of pleasure and pain moments (Kahneman et al., 2003). Kahneman and colleagues referred to the moments of experiencing pain and pleasure as *experienced utility*, which can be assessed through moment- or memory-based approaches.

Contrary to memory-based approaches that are prone to the peak-end effect (Kahneman et al., 1997), where the most intense and/or last moment of emotional responses lead to the global evaluation of an experience, moment-based assessments draw on the real-time measures of experienced utility. Experienced utility is anchored in the reality of occurring experience, not in reconstructions and evaluations of the past. According to Kahneman (Kahneman et al., 1999), the concept of well-being to which UX is strongly related (Desmet and Pohlmeyer, 2013) is constructed from the same building blocks as experienced utility, *moments*. Therefore, well-being is tightly linked to moment-based emotional experiences. Assuming this relationship, emotional experiences, when properly operationalised and measured, can account for and predict well-being.

Finally, in the new era of computing, digital technologies afford multiple forms of experiences in everyday life such as working, learning and communicating. Increasing attention has been shifted to understand how new technologies can be used to promote well-being to which positive user experience can contribute. To develop such an understanding, the foremost and critical step is to explore the nature of emotional experiences and how they can reliably be measured (Calvo and Peters, 2014).

As viewed from this perspective, exploring methodologies for assessing emotional experiences gains more weight in the HCI community.

1.1.1 Research Questions

Given the importance of UX and the necessity for measurement tools akin to UX properties (subjective, context-dependent and dynamic), the main research goal of this thesis is to identify to what extent automatic recognition of emotional expressions elicited during thinking aloud can be utilised as an approach for evaluating UX.
Prior to the main research goal, the following research questions (RQ) must be investigated:

- RQ1: Whether changes as expressed through think aloud verbalisations are sufficient enough to be measured/detected to indicate the corresponding changes in the quality of interaction with a service/product?
- RQ2: How can automatic techniques support the recognition of emotional expressions during thinking aloud?
- RQ3: What are the requirements of building an automatic system dedicated to recognition of emotions from thinking aloud utterances?
- RQ4: How can an automatic system facilitate the demanding process of analysing UX data?

Moreover, in accord with Kahneman's moment-based framework of measuring experience (Kahneman et al., 2003), the following research question has been also formulated:

- RQ5: How are users' moment-by-moment emotions elicited when interacting with a service/product integrated into their aggregated emotion and thereby attitude towards the service/product?

1.2 Approach

Automatic recognition of emotions refers to the process of analysing (such as detection and interpretation of) users' emotions from their observable cues manifested in face, voice, posture, physiological reactions, etc. by adopting automatic techniques. The overall process constitutes developing classification models that have been trained and tested on a large set of deliberately elicited emotional expressions. Therefore, having enough labelled data of emotional expressions is a first prerequisite in designing such systems. In recording emotional datasets, while there are different approaches for describing emotional expressions (for example via categorical or dimensional labels), the type of emotion elicitation methods (such as the emotions portrayed are acted or spontaneous) can vary. However, spontaneous expressed emotions are more preferable due to the fact that they better represent real-life occurrences of emotions. Identifying the existing emotional datasets for the purposes of the research motivated a systematic literature review of such corpora (Section 3.1).

With regard to the approaches used for describing emotions, data with categorical labels has the drawback of being limited to a number of emotion categories such as the basic emotions (e.g. anger, joy and surprise) or being subjected to inconsistent labelling, whereas data described with dimensional emotions (such as Pleasure, Arousal and Dominance) can be used to subsume a variety of emotional expressions even with subtle nuances in between. To tackle the limitation of data tagged with the basic categorical emotions, one approach is to convert the categories into dimensions (Section 3.4). However, both emotion models have respective strengths (and weaknesses) so that applying them together can lead to complementary information about emotional experiences (Chapter 4).

To perform automatic detection on emotions, ML techniques are used to generate classifiers based on a set of features extracted from the training input signal (Figure 4.4). While classification results based on intra-corpus experiments cannot accurately represent the performance of classifiers on unknown data, cross-corpus evaluations could better reflect the robustness and recognition performance of the classifiers. Although achieving a high recognition accuracy is a challenging task, adopting approaches such as classifier adaptation or decision-level fusion strategies could improve the recognition performance (Chapter 3).

In general, emotion recognition systems are characterised according to the modalities or channels (such as voice or face) that are used as input data. Different modalities could be utilised considering a number of factors such as the validity and reliability of the signa in real-life application settings, the cost and the intrusiveness imposed on users (Calvo and D'Mello, 2010). Hence, an automatic recognition system should be built upon the modalities that can achieve an effective combination of those factors which correspond to the purpose of the application.

Vocal and facial expressions are two of the most valid sources of information for emotional analysis and each has a long history in psychology (Ekman and Rosenberg, 1997; Scherer, 1986; Williams and Stevens, 1972). Given the proposed use of the think aloud protocol in this thesis, vocal expressions in verbalisations could be analysed against emotions without introducing additional cost and intrusiveness for users. Therefore, voice as an integrated modality to thinking aloud appears to be a plausible source for capturing data incorporated in an automatic recognition system dedicated to this protocol.

Moreover, integration of multiple modalities has the advantage of deriving complementary information about emotions (Ringeval et al., 2015). Hence, facial reactions evoked during thinking aloud could be analysed in addition to vocal expressions, especially since emotion recognition from facial expressions has shown a relatively higher accuracy rate than from vocal expressions (Calvo and D'Mello, 2010). Consequently, a more comprehensive understanding of emotional experiences could be achieved through a combination of different modalities (Chapter 5).

To evaluate the performance of an emotion recogniser, one approach is to manually annotate the emotional content of the segments and compare them against the automatic evaluations. Manual coding is a tedious task and in particular requires a great amount of time and trained coders with respect to facial expressions. Consequently given the resource demands related to the manual analysis of facial expressions, in this thesis manual coding is only applied for evaluation of vocal expressions (Sections 4.2.6 and 5.3.1).

Online shopping

To investigate the proposed methodology as an approach for evaluating emotional experiences, the online shopping domain was chosen as the scope of this research. According to the Forrester research, e-commerce generated 112 billion Euros in sales for European retailers in 2012, and is expected to yield more than $191 billion Euros in 2018 (Forrester, 2017). These numbers imply the rapid growth of e-commerce technology and its impact on the economy. At the same time, the success of web-based retailing hinges crucially on customers' emotions (Mehrabian and Russell, 1974), where positive emotions stimulate the intention to purchase and negative emotions most likely result in avoiding a product purchase (Desmet, 2012). In other words, the success of the services is to a large extent dependent on the quality of experiences that they promote. Aligning with the theory of the measurement (Hand, 2004), the quality of experience can be improved when it is appropriately measured and interpreted.

1.3 Main Contributions

The contributions of my PhD research project can be summarised as follows:

1. Previous research has endorsed multi-method, multi-operation and in-situ-based measurement approaches for assessing emotional experiences (Law and van Schaik, 2010). However, the implementation of such approaches has been regarded as impractical, due to their resource-demanding nature (Law and van Schaik, 2010). To address these hinderances, a cost-effective assessment approach is proposed whereby automated emotion recognition techniques are adopted. The application and feasibility of this approach has been investigated in two different empirical studies (Chapters 4 and 5).

2. The literature is rich in both UX related research and ML techniques, but sparse when it comes to the possible utility of ML techniques for studying UX. The overlapping realm of UX and ML research fields is identified, where ML techniques can be used for recognition of emotional expressions. Specifically, in this research, ML techniques are adopted for analysing vocal emotional expressions evoked during system interaction. To the best of my knowledge, no previous research has investigated and operationalised such source of data for understanding emotions elicited in UX.

   To apply ML techniques in the process of recognition of emotional expressions, the contributions of this research at the finer level are as follows (which are presented in Chapter 3):

   (a) A list of speech emotional corpora that are publicly available is provided. Although a number of emotional corpora has been proliferated, the availability and accessibility of such data for public use remained unclear.

   (b) An adaptive approach based on concatenating neural instances from testing data to training data is proposed and investigated in order to improve the accuracy of the automatic recognition.

3. The overall architecture of an automatic UX assessment tool built based on the think aloud protocol is proposed. Concluded from the findings of the two empirical studies, the requirements for developing such assessment tool are presented (Section 6.4).

1.4 Outline of the Thesis

Chapter 2: The definition of UX remains controversial. Thus, it is important to begin with a clear explanation of UX, the related structural models and frameworks of this notion. Exploring the extensive methodological approaches used for measuring UX, the definition and history of the think aloud protocol in usability and

UX research is elaborated. Finally, definitions of emotions and how they can be assessed unobtrusively by automated approaches during thinking aloud test sessions are given, providing the basis for the research in subsequent chapters.

Chapter 3: To assess emotions by machines, the overall architecture of the recognition systems based on the nonverbal analysis of vocal expressions is described. In particular, the main components of such systems in terms of the approaches for feature extraction from raw data, algorithms for training and finally methods for optimisation and classification are explored. The results are two emotion recognition approaches, where the first approach is based on using one English corpus for training machines, and classification on discrete and dimensional schemes of emotions. The second approach is built upon using multiple English corpora for training machines and an adaptive approach for classifying binary dimensional emotions.

Chapter 4: This chapter presents an empirical study conducted to test the utility of a voice-emotion recognition system based on the discrete and dimensional scheme of emotions. The study adopts a moment-based approach hinging on the concurrent think aloud technique and emotional analysis of vocal cues extracted from think aloud utterances. The results shed light on the utility of the think aloud technique for assessing elicited emotions. This chapter has been published and presented at the DIS 2017 conference (Soleimani and Law, 2017).

Chapter 5: A second study was conducted to test the utility of a different voice-emotion recognition approach incorporating multiple training corpora and binary classification of dimensional model of emotions. In addition to assess the performance of the recognition system introduced for usability evaluation, other confounding factors such as the situation of the use of interactive systems are investigated. Different instruments and methods are compared to validate the findings of this study. Moreover, this chapter presents its preliminary results on the application of facial expressions using the Kinect sensor as a substitute for self-reports of emotions. A part of this chapter (the effect of situational usage on emotional experiences) has been published and presented at the INTERACT 2015 conference (Soleimani and Law, 2015).

Chapter 6: Based on the findings presented in previous chapters general implications are drawn on three areas: Think aloud, UX and vocal emotional analysis. In each area, details of choices made and the corresponding alternatives for further improvements are discussed. Finally the research questions are revisited, limitations of this work are discussed and an outlook of future work is provided.

Chapter 2

Background

This chapter provides the theoretical background of UX, which positions UX as a sub-category of experience that is created and shaped through technology (Section 2.1). In addition, the existing frameworks of this notion will be described. In Section (2.1.2), the definition of usability and its relation to UX is explained. Section 2.2 looks into the UX measurement approaches, their shortcomings and strengths. The think aloud protocol, as an evaluation method is discussed in details. In addition, the utility of this technique for understanding both users' pragmatic and hedonic needs is explained. Next, the definition of emotions, as the key construct of UX is given based on the appraisal theory and the component model of emotions (Section 2.3). Furthermore, the background of automatic approaches for measuring expressive emotions such as vocal and facial is provided. At the end, section 2.5 reviews studies with respect to emotional responses in online shopping environments.

2.1 Definition of User Experience

Research on UX started at the turn of the 21st century when the HCI community realised that usability factors are narrow because they focus on the effectiveness of task completion while neglecting the associated emotional responses such as enjoyment. Hence, the concepts such as beauty (Hassenzahl, 2004), engineering joy (Hassenzahl et al., 2001), hedonic quality (Hassenzahl, 2003), pleasurable design (Jordan, 2002) drew increasing attention from the research community. The basic reason for considering the non-instrumental (such as joy and beauty) qualities originates from the humanistic view that hedonism is fundamental to human life (Hassenzahl et al., 2001). Hedonic qualities echo general human needs such as the need for communication, stimulation and novelty (Hassenzahl et al., 2001). Consequently, the focus of designing interactive products was shifted to the holistic perspective of interaction, which takes into account human's emotions and needs rather than merely the products' functionality.

The shift of emphasis from instrumental to non-instrumental qualities has reflected in the user experience definitions described by a number of authors (See (Allam and Dahlan, 2013; Law et al., 2009) for reviews). However, despite the attempt to develop a clear and well-understood definition of user experience, some of the standard definitions, such as the one elaborated by International Organisation Standardisation (ISO) 9241-210:2010 (DIS, 2009), have used fuzzy terms that themselves require further clarifications (Law et al., 2009).

Given the diverse and sometimes not clear views of user experience, it is essential for this thesis to start with laying out a comprehensive as well as simplified understanding of this notion. The scope, different frameworks and measurement methods of user experience will be discussed in the subsequent sections.

To put in simple words, user experience is a sub-category of experience that is created and shaped through technology (Hassenzahl, 2013). In other words the experiences in relation to interactive systems, products and services are known as user experience.

However, experience itself is not a straightforward concept to grasp. Experience emerges from reflection on an encountered event and constitutes feelings, thoughts and actions (Hassenzahl, 2013). Experience can be described as a continuous self-talk (or self-narration) during the moments of consciousness (Forlizzi and Battarbee, 2004), which comprises constant pain and pleasure feelings with different intensities (Hassenzahl, 2008). At any given time, an experience is formed based on one's evaluating their goals related to people, products and in general the surrounding environment (Forlizzi and Battarbee, 2004). Overall, an experience is consensually described as a subjective, situated, holistic and dynamic phenomenon (Hassenzahl, 2013).

Furthermore, the essential components of any human experience are emotions (Forlizzi and Battarbee, 2004). According to Forlizzi and Batterbee (Forlizzi and Battarbee, 2004), emotions have three basic roles in how we dispose to act: Emotions influence our plans and intentions, they regulate the procedures related to the plans, and finally they help us to evaluate the outcome. The plans are the mental representation of actions that we intend to take to fulfil our needs. These plans can be short or long ranging. They tend to change along with experienced emotions and the constant re-evaluation of a specific situation. Next, the actions we take are in coordination with our emotions to achieve our goals. The good or bad emotions change our physiological state to enable us to compare different experiences, which can result in continuing or stopping an experience. Finally, emotions enable us to evaluate the outcome of an experience and shape our future behaviour. If through the experience our goals are accomplished, the experience is evaluated as satisfactory and as a source of pleasure. On the other hand, if our plans and following activities are interrupted, negative emotions will be resulted. In such cases, emotions guide us to devise a new plan (Forlizzi and Battarbee, 2004).

The perspective of experience reviewed above can be applied for user experience resulting from interacting with products/services with emotions being the integral element. Emotions direct the way we plan to interact with products, the procedure of the actual interaction, and the appraisal and outcomes out of those interactions (Forlizzi and Battarbee, 2004).

Additionally, the experience of an interactive product/service is perceived along with two different dimensions, pragmatic and hedonic qualities of a product/service. Pragmatic quality refers to the ability of a product to realise a functionality, it focuses on the usability and utility in relation to performing a task such as "making a phone call". Hedonic quality in contrast refers to those features of a product that support non-instrumental human needs such as the desire for novelty, change, self-expression, aesthetics and stimulation. Hassenzahl 2008 has described pragmatic quality as "do-goals" and hedonic quality as "be-goals". User experience is about fulfilling be-goals through technology use. If the be-goals are experienced, hedonic emotions are attached to the product and therefore a positive user experience is achieved.

Moreover, given the role of the social context of UX, the concept of "co-experience" was introduced by Batterbee and Koskinen (Battarbee and Koskinen, 2005). Co-experience is about experiences that are created and shared between people in social interaction. However, this view of UX has been opposed by some researchers due to this reason that, although people may influence each other's experiences, yet experience is constructed from inside individuals and through the building blocks of pain and pleasure moments (Law et al., 2009).

In summary, user experience (UX) is consensually viewed through the following characteristics (Law et al., 2009):

- UX is a subjective interpretation of using interactive products/services as opposed to the objective usability metrics (e.g. number of errors).
- UX focuses on situational and dynamic aspect of an interaction. In other words, UX is context-dependent and changes over time.
- UX is interested in user's emotions as antecedent, consequence and ongoing feedback of product/service interaction.

- UX is a holistic phenomenon, which includes all aspects of the experience.

2.1.1 Existing Frameworks and Models

The existing models of UX have described this notion based on its key components and their functional interrelations. All the models have shared the agreement that the key components of UX are the aspects beyond usability. In this section, a brief review of the most referenced frameworks is provided in the following. Note that the frameworks are presented chronologically in terms of year of publication.

Forlizzi and Ford (2000) introduced a framework based on four concepts to understand the quality of an experience: Sub-consciousness, cognition, narrative, and storytelling. Sub-conscious experiences are those activities that are do not compete for one's attention and thinking process. These experiences are rather formed "thoughtlessly", such as working with products that once they are learnt, they can be used routinely. Cognition is used to represent experiences that require people's cognition or thinking processes, for instance, interacting with unfamiliar systems or environments, which demand attention and cognitive effort. Narrative experiences are those that have been formalised as language in individual's head. For example, a set of features and affordances of a product offers such a narrative of use. The storytelling concept represents the subjective aspect of the experience, where a user makes a subjective story based on prior experience, her/his emotions and the context of use.
Furthermore, shifts can occur between the components of this framework over time; For instance, a sub-conscious experience can migrate to a storytelling experience, when some levels of meaning is added and the experience is communicated to other people.

Hassenzahl (2003) proposed a model for UX with respect to four key elements (as it has been illustrated in Figure 2.1):

1. Subjective nature of experience
2. Perception of a product's qualities (apparent character), which can be understood along two general categories hedonic and pragmatic attributes
3. Emotional responses triggered as reactions to the product's attributes
4. The context or situation of use

This model explains that a product is designed as a combination of certain features such as content, presentational style, functionality and interactional style. Individu-

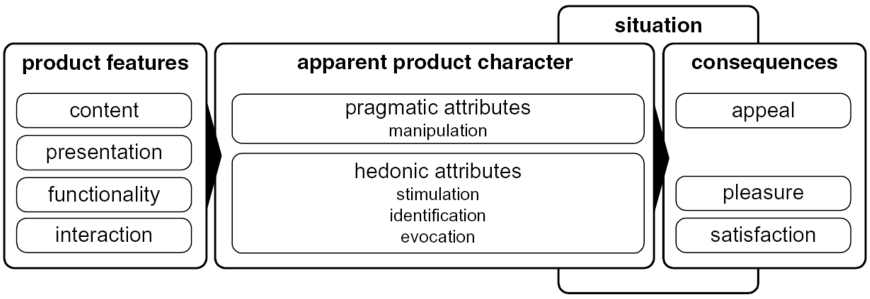


Figure 2.1: Key elements of the model of user experience (Hassenzahl, 2003).

als perceive the product's features and construct a version of the product character based on their personal standards and expectations. The character can be distinguished along two distinct instrumental/pragmatic and non-instrumental/hedonic attributes. In this regard, instrumental qualities form a product's potential to support performing a task, thus they are regarded as the quality in use. In contrast, hedonic qualities correspond to fulfilling a need, where among them, stimulation, relatedness and competence are the most salient needs to shape a positive experience (Hassenzahl, 2008). As a consequence, the perception of character in a particular situation elicits emotional responses (such as pleasure and satisfaction) or as cognitive judgments of appealingness (e.g. how good/bad a product is). Moreover, satisfaction is a consequence of fulfilment of expectations, whereas pleasure is linked to something desirable but unexpected.

The framework of sense-making by McCarthy (2004) described UX based on four interconnected *threads* of experience: Sensual, emotional, compositional, and patio-temporal strands. The sensual thread is concerned with visceral sense of the experience and the sensory engagement of the situation. Therefore the sensual thread connects a user with the experience through interaction. The emotional thread is a user's reactions and appraisal of the experience. The compositional thread refers to the narrative structure of the experience. It includes possible actions that can take place within an experience, consequences and explanations of the actions. And finally the fourth is the spatial-temporal thread, pertaining to the situation or context of the experience. Therefore, the experience is shaped by the quality of time and space. The six processes of sense-making- anticipating, connecting, interpreting, reflecting and appropriating- contribute to holistic "felt experience".

The CUE-Model (Components of User Experience) of UX proposed by Mahlke and Turing (2007), explains that experience is gained when a user with specific characteristics (such as skill and knowledge) interacts with a system to accomplish a task. As a result, the interaction is shaped within system properties and user characteristics

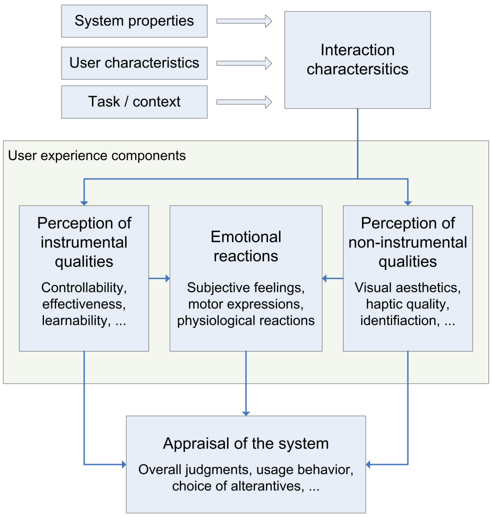

System properties
User characteristics
Task / context
Interaction charactersitics
User experience components
Perception of instrumental qualities
Controllability, effectiveness, learnability, ...
Emotional reactions
Subjective feelings, motor expressions, physiological reactions
Perception of non-instrumental qualities
Visual aesthetics, haptic quality, identifiaction, ...
Appraisal of the system
Overall judgments, usage behavior, choice of alterantives, ...

Figure 2.2: The CUE-Model: Components of User Experience (Mahlke and Thüring, 2007).

and affected by the context of use. This model assumes three major components of UX. These components are perception of instrumental and non-instrumental qualities and the subsequent emotional reactions. Here usability and usefulness of the product/service are considered as instrumental qualities, whereas appeal and attractiveness of the system (look and feel) fall into the non-instrumental qualities. Both types of qualities most likely evoke users' emotional reactions, the third component of this model. Finally, all the components together determine a user's overall evaluation of the system (Figure 2.2).

More recently, Porat and Tractinsky (2012) presented a framework grounded on the Stimulus-Organism-Response (S-O-R) model of Mehrabian and Russell (1974). According to the S-O-R model (Figure 2.3), environmental characteristics induce individuals' emotional responses. The premise of this model is that the atmospheric cues are the *Stimuli* that influence peoples cognitive and emotional reactions *Organism*, which subsequently result in certain behaviours such as to approach or avoid the environment *Response*.
Porat and Tractinsky adopted the S-O-R model for interactive products/services, where the environment (interactive applications) is perceived through its aesthetics and usability features. Likewise, the perceived characteristics of the system triggers emotional responses, which can be described through three dimensions: pleasure, arousal and dominance. Finally, the emotional responses lead to an attitude (approach or avoidance) towards the application.

A number of research studies have investigated UX focusing on a few of the constructs of this notion and their interrelationships rather than UX as a whole. For

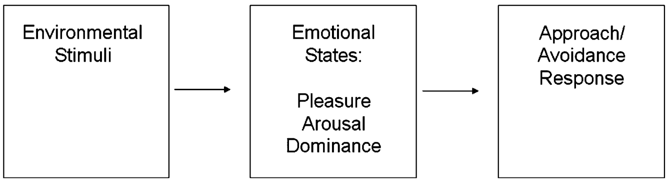


Figure 2.3: The Environmental Psychology Model (Mehrabian and Russell, 1974).

example, Tuch and colleagues (2012) analysed the relationship between two constructs of UX, usability and aesthetics. In their study it was observed that good usability improves the perceived aesthetics and emotion was found to be the mediating factor in this relationship. Another study (Gross and Bongartz, 2012), explored UX based on the product-specificity aspect, where the type of product (being as more pragmatic or hedonic) had impact on the importance of the perceived usability, aesthetics or fun. For instance, for a product with a pragmatic purpose of use, usability quality was found to be more weighed than the aesthetic or fun qualities.

The situational aspect of UX is another scope that has been investigated in a number of studies. For example in a study conducted by Hassenzahl and colleagues (2008), it was shown that depending on the situation, individuals might form different perceptions of an experience and its overall quality. Along the same lines, Hassenzahl and colleagues (2002) tested two different situations, task- and fun- related goals in interactions with websites. Their results showed that the aspect of usability had a stronger effect on the overall evaluation in individuals with task-related goals than in individuals with fun seeking goals. In another study by Hartmann and colleagues (2007), participants were situated in two different scenarios, looking for a summer internship or researching on their PhD topic, and asked to interact with three different website designs.Their results confirmed that in a fun-related scenario (summer internship), hedonic qualities (such as the look and feel) are the most important attributes, whereas in a task-related scenario (PhD research), pragmatic qualities (such as the content of the websites) are given more weight as the important attributes.

The motivational aspect of UX has been specifically emphasised on the context of online shopping. In this regard, two types of motivational orientations have been generally reported: 1) Fun or hedonic orientation, where consumers shop because of the hedonic values such as fun, pleasure, fantasy, escapism or, in other words, the experiential goal of shopping; 2) pragmatic or utilitarian orientation, where consumers shop based on their utilitarian goals, such as to acquire items effectively and efficiently (Bui and Kemp, 2013; O'Brien, 2010).

It has been observed that the motivation of shopping moderates the relationship between arousal, pleasantness and shopping behaviours (Kaltcheva and Weitz, 2006). For instance, a high-arousal environment (e.g. warmer or high saturated colours) led to pleasure for customers with hedonic motivations (e.g. to obtain gratification out of shopping), whereas customers with pragmatic motivations had less pleasure due to their desire to minimise the energy needed to complete the activity. In a study (O'Brien, 2010), the relationships of the motivational orientation with engagement was investigated. The results showed that customers with a hedonic motivation were more engaged in an online shopping context.

2.1.2 Usability

As it was mentioned in the section 2.1, the concept of UX was constructed based on usability to encompass all aspects of interactive experiences such as pleasure and beauty. UX could be seen as an elaborated form of satisfaction - one of three prototypical metrics in usability (Law and van Schaik, 2010). Consequently, the background of UX would be incomplete without clarifying the definition of usability.

Usability is defined in ISO 9241-11:1998 as *The extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.* Usability is as an essential concept for capturing the quality of use of interactive technology, as indicated by the three prototypical metrics: effectiveness, efficiency and satisfaction. Effectiveness describes the extent of achieving users' goals in terms of accuracy and completeness with system interaction. Efficiency represents the effectiveness of system usage in regard to the consumed resources such as time and effort, and satisfaction relates to users' comfort and acceptance of the system.

The inclusion of pragmatic or instrumental component in different models of UX supports the assumption that usability is subsumed by UX (Law and van Schaik, 2010). The distinction between usability and UX lies on the broader scope of UX in terms of its compositional attributes as well as their measurability as it was discussed in Section 2.1. In addition, the partially objective perspective of usability is in contrast with the mostly subjective view of UX. While usability metrics are gauged by objective measures such as time-on-task and completion rate, a balanced use of both subjective and objective measures has been recommended for measuring UX (Law and van Schaik, 2010).

2.2 Measurement Approaches to User Experience

Methods for measuring UX are largely drawn from usability (Albert and Tullis, 2013). This again underlies the fact that the concept of UX was developed based on usability as an extension to address the limitations of the usability metrics. However, UX encompasses much more complex aspects, which are mainly associated with the psycho-physiological characteristics of experience such as emotions (McCarthy and Wright, 2004), intrinsic needs (Hassenzahl, 2003) and aesthetics qualities (Lavie and Tractinsky, 2004).

To explore the methodological approaches used for UX, Bargas and colleagues (Bargas-Avila and Hornbæk, 2011) reported the results of reviewing 51 studies from 2005-2009 and Law and colleagues (2014) summarised the results of 58 selected studies from 2010-2012. According to their reports: The methodologies used in the UX studies were mostly qualitative. Among others, questionnaires, interviews, user observations and video recordings were the frequently used collection methods. In addition to those methods, an increase use of physiological measurements such as heart rate (HR), galvanic skin response (GSR) and electroencephalogram (EEG) could be observed over the years (five studies reported by Law compared to three studies reported by Barges). This increased use of qualitative and physiological methods reflected the shift of interest from usability to UX which required to be accompanied by a change in methodology as well. Likewise, quantitative ratings have shown not to provide sufficient information especially about the emotional responses of an experience (Kujala et al., 2014).

While the self-report methods such as questionnaires and interviews provide a subjective evaluation of UX, physiological measurements can be used to assess UX objectively, tracking the changes over time (Lin et al., 2008). For example, Mandryk and colleagues (2006) investigated the changes in HR and GSR as indicators of fun and engagement with entertainment technology. Similarly, (Bruun et al., 2016a) explored the association between GSR and the state of frustration, and (O'Brien and Lebow, 2013) found links between changes in electrodermal activity (EDA) and interest in online news websites.

Despite the advantages of using physiological measurements, there are several limitations associated with them. First, taking such measurements requires wearing sensors, some of which are obtrusive (e.g., headgear for EEG sensors) and some may cause discomfort such as tension and restraint, as some sensors are highly sensitive to movement (Chandra and Calderon, 2005). Third, a robust approach enabling direct mapping of physiological data to specific emotions such as frustration is yet

to be developed (Bruun et al., 2016a). Moreover, employing physiological measurements involves additional expenses such as the purchase of special equipment and expertise that entails training (Lopatovska and Arapakis, 2011).
Behavioural metrics such as time on task, browsing time and number of clicks have been employed for assessing the relationship between motor behaviour and UX constructs such as emotion (Zimmermann et al., 2006). According to the literature (Law et al., 2009), to capture a full spectrum of UX, more than one measurement method should be employed. For example, the combination of qualitative and quantitative measures can give contextual information as well as a summary assessment of UX, or objective data collection in conjunction with the subjective evaluations can give a holistic understanding of UX (Wastell and Newman, 1996). In summary, a multi-method multi-operation measurement approach should be taken to account for different constructs of UX (Gray and Salzman, 1998). However, combination of mixed approaches implies triangulation of the different measures, which can be very costly and requires carefully planned experiment protocols (Law and van Schaik, 2010).

There are studies that have proposed other unconventional methods for measuring aspects of UX. For example, (Kujala et al., 2014) proposed the potential use of projective techniques such as sentence completion for identifying user needs, values and experiences. The use of eye tracking and facial expression analysis have been suggested in recognition of various emotional reactions such as frustration, amusement and surprise during an interaction (Bergstrom and Schall, 2014). The enhanced version of the Experience Sampling Method (ESM) by use of cameras in mobile devices and visual analysis (Larson and Csikszentmihalyi, 1983) is an alternative approach for measuring UX in situ (Niforatos et al., 2015). ESM is a research method to study people's emotions and thoughts during their daily lives. In this method, individuals are asked to self-report their momentary emotions at random occasions during different activities.

Moreover, the think aloud technique as a practical usability evaluation method has been applied in a handful number of studies to evaluate emotional experiences (e.g. Petrie and Precious, 2010; Vasalou and Bänziger, 2006). Due to the central role of this technique in this thesis, a detailed background is provided in the next section.

2.2.1 Think Aloud

The think aloud protocol has its roots in cognitive psychology. This protocol was initially described by Karl Duncker as a method to understand his participants' development of thoughts (Nielsen et al., 2002). Later, Ericsson and Simon on Protocol Analysis (1993) posited that thinking aloud is a valid data collection method to understand the process of underlying problem solving. The assumption behind this protocol is that all cognitive processes that are heeded, go through short-term memory (STM) and retained there at the time of being processed. As the result of being brought into STM, the conscious thoughts can be verbalised at the time they are perceived and shortly after the process has taken place.

Ericsson and Simon (1993) defined three levels of verbalisation, in which the degree of reliability of verbalised data decreased from the level one to the level three. In this regard, the degree of reliability corresponds to the amount of interference caused by not task-related cognitive processing. Verbalisations of this level are those that do not need to be transformed before being verbalised during performing a task. For example, verbalising a sequence of numbers by a subject during a math problem results in the level 1 of verbalisation. The reason behind this categorisation is that the numbers can be directly verbalised in the same way they were encoded in the STM and without any transformation. According to the authors, this level is the most reliable form of verbalisation.

Verbalisations of the second level are those that required to be encoded to other forms before being verbalised during task performance. For instance, images or abstract concepts must be transformed into words before they can be verbalised. As long as this transformation into words is the only intermediate cognitive process between STM and verbalisation, such verbalisations are considered as reliable data. Verbalisations of the third level are those that need additional cognitive processing beyond that required for task performance or verbalisation. For example, filtering out some thoughts during verbalisation (e.g. verbalise only information regarding a particular topic). In this kind of verbalisation, the flow of information in STM is changed during the task performance. For Ericsson and Simon, verbalisation of the level one and two can be relied upon, if properly collected. On the other hand, verbalisation of the level three of was not recommended by Ericsson and Simon's model as highly reliable for cognitive analysis.

The procedure by Ericsson and Simon defines minimum interaction between researchers and participants during task performance. Therefore, the only intervention should be the reminder to think aloud, after a period of silence (between 15 to 60 seconds). This reminder should be short and non-directive as well such as "Keep talking". Any other interactions such as neutral comments or questions lead to redirecting the participant's attention should be avoided.

Since Ericsson and Simon's model of collecting verbal data, researchers in different areas including usability testing (Nielsen, 1994b), engineering fields (Sanderson,

1990) and reading comprehension (Pressley and Afflerbach, 1995), to name just a few, showed interest in think aloud method for collecting verbal data. In the case of usability research, the main application is identifying usability problems through verbal reports and system interaction (Nielsen, 1994b). However according to a research carried out by Boren and Rarmey (2000), there are discrepancies in applying the think aloud method based on Ericsson and Simon's theory in actual usability practices. For instance, Ericsson and Simon's model describes a short and non-directive reminder “Keep talking” to prompt thinking aloud after a silence, whereas in practice, other longer forms such as “Don't forget to tell me what you're thinking?” have been usually used. Another discrepancy fell in the type of intervention, for example when the participant is stuck in a task and thus clarification is requested from the practitioner. This scenario would produce irrelevant or unreliable data under Ericsson and Simon's model. Similar scenarios could be extended to the times of technical errors or system crashes, where the usability practitioner must step in, or when the participant thinks the task has been completed when actually has not, where the practitioner's intervention is required. Consequently, the exact recommendations of Ericsson and Simon's protocol cannot always be met during usability tests.

To tackle the technical issues of applying the think aloud protocol in a usability test, an alternative approach as the “Speech Communication” was proposed by Boren and Ramey (2000). In this approach, the role of a user (a speaker), who interacts with the interface, giving information about it, and with a facilitator (a listener), who is there to actively listen to the information, have been clearly defined. According to Boren and Ramey, while the facilitator should reaffirm his/her presence by saying “yes”, “mm hm?” and “uh-huh?” at regular intervals in a test session, at the same time, she/he is required to avoid unnecessary interventions. Hereby, those types of acknowledgement communicate the active engagement of the listener/facilitator, which can promote the participant' constant speakership.

Furthermore, according to the speech communication theory, the use of token passing (such as “yes” and “mm hm”) would not have any impact on task completion due to the fact that the tokens do not introduce new content into one's short term memory, and in addition, no (or a little) effort is required for processing them. This type of thinking aloud is often called *relaxed thinking aloud*, in which users are asked to provide a commentary summary of their thoughts, what they do and their reflection on their own actions (Hertzum et al., 2015). According to Hertzum and colleagues (2015), relaxed thinking aloud is a combination of levels one to three of the classic Ericsson and Simon's model, thus including thoughts and feelings, explanations and reflections.

To understand the effects of employing the classic or relaxed thinking aloud on the value of verbal data, Zhao and colleagues (2014) compared the two versions of thinking aloud based on the value of the verbalised content in identification of usability problems. Their results suggested in favour of relaxed thinking aloud in terms of producing valuable verbal content pertinent to usability issues. On the other hand, earlier research carried out by Krahmer and Ummelen (2004) suggested no difference between the two think aloud versions in terms of the types of usability problems found. In addition, discrepancies exist between the findings in comparing the two think aloud approaches in terms of task completion time, mental workload and behavioural tendencies (Hertzum et al., 2009).

Initially, Jakob Nielsen suggested five as a sufficient number of participants needed in usability thinking aloud tests to identify the majority of problems (Nielsen, 1994a). However, research (Schmettow, 2012) has argued that factors such as the context of the study and data quality should be taken into account when estimating the required number of participants, hence the five-participants rule is not a fixed and reliable sample-size estimation for discovering problems across different studies.
In regard to investigating research questions within thinking aloud sessions (e.g. the effects of concurrent or retrospective tests on the quality of verbal data), different numbers of participants have been reported in different studies. For example, Hertzum and colleagues conducted a between group study with 14 participants to explore moderated or unmoderated users's verbalisations in relaxed thinking aloud (Hertzum et al., 2015). Zhao and McDonald (Zhao et al., 2014) recruited 16 participants to compare verbal data between the two classic and relaxed styles of thinking aloud, whereas having similar objectives, another study conducted their research by recruiting 48 participants (Bowers and Snyder, 1990).

The think aloud protocol can be conducted retrospectively or concurrently (Ericsson and Simon, 1993). For concurrent protocol, participants think aloud while they are carrying out their tasks. Alternatively, participants can first perform their tasks silently, and afterwards they verbalise their thoughts in retrospect. The latter approach is called retrospective think aloud. There are both strengths and drawbacks associated with retrospective think aloud. For example, Bowers and Snyder (1990) found that fewer verbalisations were produced in the retrospective as compared to the concurrent approach, although the retrospective protocols were explanatory in nature. In another study, it was found that there were no differences in terms of the number and type of usability problems between the two approaches. However, the concurrent protocol resulted in identifying more problems from observation, whereas

Think Aloud	Service	Participant
Concurrent vs. Retrospective (Bowers and Snyder, 1990)	Microsoft Word	48
Concurrent vs. Retrospective (Van Den Haak et al., 2003)	Online library	40
Emotional evaluation (Vasalou and Bänziger, 2006)	Yahoo! website	20
Classic vs. Relaxed (Krahmer and Ummelen, 2004)	Travel agency website	16
Concurrent vs. Retrospective (McDonald et al., 2013)	University website	10
Classic vs. Relaxed (Zhao et al., 2014)	Travel agency website	16
Moderated vs. Unmoderated (Hertzum et al., 2015)	News website of music	14

Table 2.1: Examples of previous studies with regard to number of participants and services under evaluation employing the think aloud approach.

the retrospective protocol yielded more verbalised problems (Van Den Haak et al., 2003).

Finally, think aloud sessions can be held with or without the presence of an experimenter, the former is called moderated and the latter, unmoderated. In unmoderated tests, participants are usually recruited via crowdsourcing. The crowdsourcing approach helps recruitment of individuals with diverse background, low cost and holding multiple sessions simultaneously. Moreover, findings of the related research showed that there was no difference in the content of verbalisations made by moderated and unmoderated sessions (Hertzum et al., 2015). Table 2.1 provides a summary of a few examples of thinking aloud studies with respect to their research questions, number of participants and the service under evaluation. All these studies carried out in a between group setting.

Content Analysis of Verbal Data

Transcription and segmentation of verbal data at the level of individual utterances must be undertaken prior to the content analysis of thinking aloud protocols (Nielsen et al., 2002). In this regard, careful consideration should be given to the appropriate segmentation, because this step will affect the results of content analysis. Ericsson and Simon described each segment as an instance of a general process, which could be identified by clues such as pauses and even intonations. On the other hand, according to Chi (Chi, 1997), segmentation must be determined in a way that each segment can be coded independently. In other words, the reason for segmentation is to define what constitutes a unit of analysis for coding.

The boundaries of the units can be identified based on some measures such as some language related-syntax (e.g. word and sentence) or other objective measures such as duration, turn taking and pauses (Chi, 1997). The main advantages of using the objective measures are that the time-consuming process of reading and interpreting the verbal contents can be avoided and the subjectivity of interpreting them can be mitigated; On the other hand, segmentation based on semantic features (e.g.

meaningful units) such as ideas or actions is psychologically relevant due to the fact that each segment in this way could convey certain thought and feeling (Chi, 1997). A semantically meaningful segment is called a full/complete thought (Trickett and Trafton, 2009), which is linked to a cognitive process. Subsequently as results of segmentations, the coding schemes can be created.

To analyse verbalisations, the common approach is creating a coding scheme, on which categorisation of the verbal contents based. Examples of categories include topic, valence and relevance of the content (e.g. Hertzum et al., 2015; Zhao et al., 2014). To briefly explain, for topic, the reported categories are usually: action description (the procedure of activities), reasons to performing an action in a specific manner (explanation), redesign proposal and also the general UX of the system, which describes positive or negative experiences elicited from the use of a system (e.g. Cooke, 2010; Zhao et al., 2014). For valence of utterances, the content is categorised based on their positive and negative connotation (such as "I hate all these ads" as a negative statement). Overall, positive valence is derived comments that convey the state of approval, satisfaction and positive experiential aspect of interaction. Conversely, negative valence describes the states of disapproval, dissatisfaction and negative reactions toward the system. Finally, the relevance of a comment could be defined by three levels: low, medium and high, where each level describes to what extent the given comment supports the identification of usability problem or a positive usability aspect.

Traditionally, content of verbalisations was used as the source of data for discovering usability problems. However, with technologies moving beyond task performance, usability is no longer the only concern of practitioners. The focus has been extended toward obtaining knowledge about users' hedonic and emotional experiences as well. Examples can be seen from the studies, where the content analysed not just based on the description of thoughts and actions, but also included experiential aspects such as emotions (Hertzum et al., 2015; Petrie and Precious, 2010). Earlier studies have also shown that interacting with avatars as experimenters could elicit in users emotional responses associated with their personal values and experiences (Vasalou and Bänziger, 2006).

Therefore, in line with the objectives of an evaluation (whether the objectives are identifying experiential or usability issues or both), data collected from thinking aloud should be analysed accordingly. One approach is to code the verbal content to encompass experiential criteria such as emotions as well as the performance related aspects. Previous research has shown that users are able to describe their emotive responses while they are thinking aloud (Hertzum et al., 2015). Hence, searching for words such as “exciting” and “annoying” within the verbal content could be used as a method. For example, Hertzum and Holmegaard (Hertzum et al., 2015) tested a news website by asking participants to think aloud, which entailed the explanation of their thoughts, what they like or dislike about the website. Zhao and colleagues employed the think aloud method to evaluate a travel agency website (Zhao et al., 2014) by instructing participants to to say out loud everything that they would say to themselves silently. Similarly Marki and Cox assessed digital libraries (Makri et al., 2010) by asking participants to verbalise what they were doing and any thoughts going through their mind.

Analysing data collected in think aloud sessions have been carried out until recently by manual human coding (such as Hertzum et al., 2015) and (Zhao et al., 2014)). On the other hand, automatic approaches for emotion detection from verbal and nonverbal clues such the analysis of vocal and facial expressions elicited during system interaction could expedite the process of coding (Vasalou and Bänziger, 2006). This research also embraces the automatic approaches for evaluating emotional experiences. Hence, background on the definition of emotions is essential, which will be discussed in the next section. In addition to the definition, the the use of emotion recognition approaches and their history in psychology as well as information technology are elaborated.

2.3 Definition of Emotions

The definition of emotions has been widely discussed in psychology (Kleinginna and Kleinginna, 1981). Scherer (2005) has described emotions as cognitive responses with relatively high intensity and short duration, from seconds to minutes, in relation to external or internal stimulus events. According to Scherer’s definition, emotions are appraisal driven. He has explained that emotions are triggered when the implications of an event are relevant to one’s concerns, such as needs, goals and the desire for intrinsic pleasantness and novelty. Given the importance of the event and its consequences, a person’s (or organism) subsystems (such as the Central Nervous System, Automatic Nervous System, Somatic Nervous System and etc.) are coordinated to respond appropriately to that event. These responses are the important features of emotions that can be empirically measured.

In addition, stimuli and the associated appraisal usually change rapidly due to new information or re-evaluation. As a result, the corresponding responses are also likely to change constantly for readjustment to the new circumstances. Moreover, Scherer has explained that emotions have strong effect on action tendency, generating new

motivations and plans, which result in interrupting an ongoing behaviour and initiating new actions.
Considering the above characteristic, the Component Model of Emotions (CPM) by Scherer has described emotions through the process of five interrelated components:

- Cognitive component: appraisal of events and objects
- Neurophysiological component: regulation of the bodily system which is manifested in physiological changes such as sweat production
- Motivational component: preparation and direction of action based on the motivational effect of events
- Motor expression component: communication of reaction and behavioural intention represented in facial and vocal expression
- Subjective feeling component: the subjective experience of emotional state once it has occurred

Emotions differ from moods, which have longer duration (hours or even days) and emerge without any apparent cause and are not necessarily linked to a specific event. Emotions have been distinguished from core affect (Russell, 2003) as well (Scherer, 2005). In this sense, affective states are behavioural tendencies of a person, which make her/him more inclined to experience certain moods more often or to react with certain types of emotions.
While there are ongoing debates on the differences between affect, emotion and mood (Ekkekakis, 2013; Wetherell, 2012), I do not delve into details as it entails a separate exposition. The terms emotional state and affective state were used interchangeably in this thesis. Both terms have been used to account for the dynamic state, when a person experiences an emotion.

2.3.1 Descriptive Scheme of Emotions

Emotions have been described predominantly based on two models. The model of basic emotions posits that there is a set of discrete emotions with specific and universal patterns in facial expressions (Ekman, 1992; Izard, 2013). In the relevant theories, the number and nature of the basic emotions vary. However, anger, joy, sadness, fear, surprise and disgust are generally included. Other emotions (such as contempt) are considered as blends of basic emotions (Plutchik, 1980).
The other common model has described emotions in terms of continuous dimensions. The most common dimensions for emotional experience are valence (pleasure) and

arousal (activation) (Mehu and Scherer, 2015; Russell, 1980). Valence describes the positive or negative value of an emotional states and arousal describes the strength of the individual's disposition to act. The third dimension that also has also been used is dominance (power or control), which describes the individual's power to deal with relevant events. The three dimensional model of emotion, valence (pleasure), arousal and dominance (PAD) was proposed by Mehrabian and Russell (1974). Mehrabian (1996) demonstrated that any type of emotion can be represented by the nearly independent dimensions of PAD. Since then, the PAD model was applied for subjective evaluation of emotions in various studies (e.g. Eroglu et al., 2001; Hsieh et al., 2014).

To summarise, this thesis adopts the notion of emotion defined by Scherer (2005) as involving cognitive and evaluative elements. Furthermore, to describe emotions, I considered both discrete and dimensional schemes for describing and evaluating emotions in the context of UX.

2.3.2 Emotions in User Experience

There is a bulk of research examining the relations between emotions and the other constructs of UX (e.g., aesthetic quality, user needs). To assess actual experience and emotion, the adopted approach can be carried out concurrently and/or retrospectively by the means of different methods such as self-report questionnaire, physiological measurements and automatic recognition approaches.

In a series of studies conducted by Redelmeier and Kahneman (1996), patients were asked to continuously report the intensity of pain experienced during the procedure. After the procedure, they were also asked to provide a retrospective judgment about the overall level of pain during the procedure. It was found that, the retrospective judgments were the indicators of the most intense (peak) and/or the end moments of the experienced pain (*peak-end rule*). Similar results were observed in different situations, such as the emotional experiences of watching a movie (Fredrickson and Kahneman, 1993). Based on these findings, the researchers adopted the approaches to carry out similar studies in the field of HCI. In doing so, an interactive experience is usually split into a sequence of subtasks, where after each subtask, participants are asked to complete a survey.
For example Hassenzahl and colleagues (2007) explored the relation between mental effort and emotional valence during an interaction. Results of their study revealed a negative correlation between mental effort and valence, although both aspects were positively correlated with retrospective affinity for the experience. In another

study, two systems, one with a higher number of usability problems, were examined using self-reported ratings and GSR sensors (Bruun and Ahm, 2015). While GSR sensors showed a higher fluctuation for the system with more usability problems, yet retrospective self-reports of emotions showed a better rating than the corresponding concurrent ratings for the system with more usability problems.

On the other hand, retrospective self-reports have been predominantly used in previous research. Examples are investigating the relations between intrinsic needs (such as competence and relatedness using a self-developed questionnaire) and emotions using the PANAS questionnaire in (Hassenzahl, 2008), between pleasurable stimulation (as a hedonic value) and emotional arousal by taking SAM scales (Hassenzahl et al., 2008) and between perceived aesthetic as a hedonic attribute and emotional pleasure and arousal employing the Porat and Tractinsky scale (Porat and Tractinsky, 2012).

Moreover, previous work has emphasised measuring emotions through a combination of methodological approaches. For example, Mahlke and Minge (2008) presented an HCI related experiment based on Scherer's Component Process Model of emotions (CPM), where emotions are responses from five interrelated organismic subsystems (component) when evaluating an external event. Subsequently, the authors proposed that these responses could be measured using methods corresponding to each of the components, including subjective ratings, physiological measurements, facial expressions, behavioural tendencies tracking, and assessing cognitive appraisal through questionnaires (examples of applying different assessment methods for measuring each component of emotions are given in Table 2.2). However, in Mahlke and Minge (2008) study, significant correlations were only confirmed between subjective emotions and measures obtained from the other four components. In other words, no correlations were found amongst the other components of emotions such as physiological changes, expressive reactions and behavioural tendencies. Therefore, it implies that subjective evaluations could be used as indicators of the other components of emotions. Conversely, measurements of the other components of emotions could reveal the degree of subjective ratings.

As it was mentioned earlier, traditional approaches prevalently relied on the use of well-defined self-report scales. The following reviews the most validated and widely used instruments that have been employed in this research as well:

Self-assessment Manikin: SAM is a scale for measuring emotions based on the dimensional scheme (Bradley and Lang, 1994). The underlying assumption in using

Components of Emotions	Measurement methods
Subjective feelings	SAM (Bruun et al., 2016b; Mahlke and Minge, 2008)
Physiological reactions	EDA and Heart rate (Mahlke and Minge, 2008), GSR (Bruun et al., 2016b)
Motor expressions	EMG (Mahlke and Minge, 2008), Video-based facial analysis (Staiano et al., 2012)
Cognitive appraisal	Geneva appraisal questionnaire and Retrospective think aloud (Mahlke and Minge, 2008)
Behavioural tendencies	Time for task completion (Mahlke and Minge, 2008), Mouse click and Mouse movement (Zimmermann et al., 2006), Rhythm of typing on keyboard (Epp et al., 2011)

Table 2.2: Examples of applying assessment methods for measuring each component of emotions in corresponding studies.

SAM is that individuals are the best source of information on their emotional experiences (Mahlke, 2008). SAM consists of pictures of manikins in a scale for each of the dimensions: valence, arousal and dominance. SAM scales range from a smiling, happy figure to a frowning, unhappy figure for valence and from a wide-awake, excited figure to sleepy, relaxed figure for arousal. The dominance dimension in SAM represents control with changes in size of the manikins, where a large figure implies maximum control in the situation (Figure 2.4).

Semantic Differential Scale: This scale was developed by Mehrabian and Russell Mehrabian and Russell (1974) and has described emotions along with the PAD framework. The Semantic Differential Scale consists of bipolar emotion labels for measuring valence (pleasure) (5 items), arousal (6 items) and dominance (6 items) (Table 2.3) along a 9-point scale. The bipolarity nature of each dimension and their independence have been confirmed in various studies (e.g. Feldman Barrett and Russell, 1998; Mehrabian, 1995).

Differential Emotion Scale: The DES is another instrument for measuring emotions (Izard, 1993). The DES scale measures ten fundamental emotions relying on this theoretical background that these ten basic emotions are universal and distinguishable in human facial expressions. The vocabularies in the DES items were obtained from the research on the verbal labels of facial expressions (Izard, 1993). The DES items are thirty emotional adjectives to assess ten emotions (three adjectives for each emotion): Joy, interest, surprise, sadness, anger, disgust, contempt, fear, shame and guilt. Each adjective is administered on a 5-point Likert scale from not at all to very strongly.

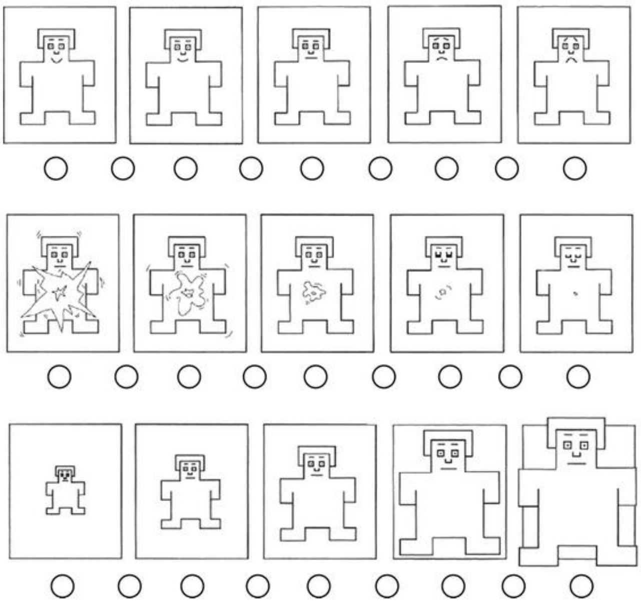


Figure 2.4: SAMs used to rate the dimensions of valence (top panel), arousal (middle panel), and dominance (bottom panel).

Dimension	Items
Pleasure	Happy–unhappy Satisfied–disappointed Hopeful–despairing Relaxed–bored Content–annoyed
Arousal	Energetic–languid Excited–calm Wide awake–sleepy Restless–slow Aroused–unaroused Enthusiastic–serene
Dominance	In control–helpless Respected–insignificant Dominant–submissive Autonomous–guided Active–passive Free–restricted

Table 2.3: Bipolar items for measuring the Pleasure, Arousal and Dominance dimensions in the Porat and Tractinsky Scale.

2.4 Expressive Emotions

2.4.1 Vocal Expression of Emotions

Human listeners are good at inferring emotional states from vocal expressions. This reliable evidence implies that voice is a significant carrier of affective signals (Banse and Scherer, 1996). In addition, results of research studies have confirmed that human listeners are able to reliably discriminate among emotional states on the basis of vocal cues (Banse and Scherer, 1996; Pittam and Scherer, 1993). This means that acoustic features are differently patterned for communicating different emotions (Banse and Scherer, 1996). Consequently, specific acoustic characteristics of the concurrent vocalisations are used to convey different emotional states.

Research on emotional cues manifested in voices has a history of about 45 years, starting with the work of Williams and Steven in 1972 (Williams and Stevens, 1972). The theoretical idea behind vocal emotion analysis is that affective states have a specific effect on the somatic nervous system (SNS), which in turn changes the tension of musculature and thus influences the voice production. In other words, the characteristics of the vocal expressions at a particular time describe the effect of emotional changes on the SNS (Scherer, 1986). Acoustic features such as a) pitch: the level, range and contour of the fundamental frequency, which is referred as F0, b) intensity or the vocal energy of the voice and c) duration are strongly involved in vocal emotion expression. These features are referred as prosodic features, which appear to convey the flow and rhythm of speech (Koolagudi and Rao, 2012). In other words, the prosodic features represent the perceptual properties of speech (Rao and Yegnanarayana, 2006).

Previous research has shown the contribution of prosodic features in different emotional expressions as summarised by Pittam and Scherer (1993): Anger is generally associated with an increase in mean F0 and mean energy. Fear can be identified with increase in mean F0, increase in F0 range and high frequency energy. Sadness is usually characterised with decrease in mean F0, F0 range and mean energy. For joy, a strong association has been found with increase in mean F0, F0 range, F0 variability and mean energy.

In addition to the prosodic features, phonetic features such as Mel-Frequency Cepstral Coefficient (MFCC), a feature introduced by Davis and Mermelstein (1980), which represents vocal tract information, and the voice quality features such as jitter and harmonics-to-noise ratio (HNR) have also been frequently used for identifying emotions (Sato and Obuchi, 2007). The Fourier transform of a speech frame gives a short time spectrum. Features such as spectral energy and slope and formants can be obtained from a spectrum. Finally the Fourier transform on log magnitude

Emotion	Pitch mean	Pitch range
Anger	Increased	Wider
Happiness	Increased	Wider
Sadness	Decreased	Narrower
Surprise	Normal or increased	Wider
Disgust	Decreased	Wider or narrower
Fear	Increased or decreased	Wider or narrower

Table 2.4: Correlations between emotions and pitch-related features by applying the statistical functions mean and range (Morrison et al., 2007).

spectrum results in cepstrum of a speech frame. These vocal tract features are sometimes referred to as spectral features. In general, the combination of prosodic and spectral features can improve the emotion recognition performance (Koolagudi and Rao, 2012).

To infer emotions from a given speech signal, one common approach is computing the descriptive statistical information (such as the mean, range, variance, maximum and minimum) of the LLDs at the sentence level. This is done by initially extracting the acoustic features at a fix frame rate (e.g. 5ms or 10ms). Next the extracted features are summarised into a single value by applying the statistical functions over the given sentence. For example Table 2.4 shows the correlations between emotions and the pitch feature when the statistical functions (e.g. mean and range) are applied (adopted from Morrison et al., 2007).

Given the correlations between emotions and acoustic features, studies such as the one conducted by Nowak and colleagues (2012) used Praat (Boersma et al., 2002), the phonetics software, to extract manually a small number of acoustic attributes for emotion recognition. Alternatively, the use of predefined feature sets, which contain a large number of emotion-related acoustic attributes, has been increasingly adopted for automatic emotion recognition. The study examples include voice emotion detection in negotiation systems to visualise feedback (Nowak et al., 2012), tutoring applications to improve communication (Pfister and Robinson, 2011), entertainment and game programs (Jones and Deeming, 2008), game applications for children with autism (Marchi et al., 2015b), emotionally-intelligent car systems (Schuller et al., 2007c), emotion tracking applications integrated in virtual agents (Vogt et al., 2008) and vocal chat media (Dai et al., 2015).

There are a number of popular and predefined feature sets that have been developed and tested based on previous studies on voice emotion analysis (Batliner et al., 2011; Schuller et al., 2003):

- The INTERSPEECH 2009 Emotion Challenge (EC) Set (Schuller et al., 2009c) is the first baseline feature set for emotion recognition. The INTERSPEECH 2009 includes two sets; the standard set with 384 attributes and the extended set containing 6552 acoustic attributes.

- The INTERSPEECH 2010 Paralinguistic Challenge Set with 1582 acoustic attributes, which was designed for age recognition as well as for level of interest (Schuller et al., 2010a).

- The INTERSPEECH 2011 Speaker State Challenge Set for identifying speaker states (e.g. state of sleepiness), consisting of 4368 features (Schuller et al., 2011b).

- The INTERSPEECH 2012 Speaker Trait Challenge Set with 6125 features, which was developed for recognition on speaker traits (such as the level of openness and consciousness, Schuller et al., 2012).

- GeMAPS: In contrast to the large scale feature sets, smaller standard parameter sets such as the minimalistic Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and its extended version (eGeMAPS) consisting of 62 and 88 parameters respectively (Eyben et al., 2016), have been proposed recently for emotion recognition. The acoustic parameters of these sets were selected based on their potential value on indexing emotions, their automatic extractability and their theoretical significance based on recommendations of former research (Eyben et al., 2016).

Overall, the presented sets vary in the number and type of acoustic features contained as well as the functions applied onto the features (see Eyben, 2015, for a list of standard baseline feature sets and detailed information of the parameters in each set).

After feature extraction, ML classifiers can be used for recognising emotions from the vector features. Implementations of these classifiers have been provided in different toolboxes such as WEKA (Hall et al., 2009). Classifiers vary in terms of their generalisability and ability to handle high dimensionality and non-linear problems (i.e. their recognition rate in different applications). Other considerations when choosing a classifier are efficiency such as real-time capability, low computational and memory cost (Schuller et al., 2011a).

Popular classifiers for emotion recognition have been K-Nearest Neighbour (KNN) classifiers and Linear Discriminant Classifiers (LDCs) that have been successfully used since the very first studies (e.g Jones and Deeming, 2008). These classifiers also have shown a good recognition rate for non-acted emotional speech (Kwon et al., 2003). LDCs have been extended to become Support Vector Machines (SVM) (Cortes and Vapnik, 1995), which create straight-line decision surfaces (hyperplanes) to discriminate between classes.

SVM classifiers provide good generalisation properties independent of the dimension of feature vectors (Kudoh and Matsumoto, 2000). On the other hand, in linear SVMs, the hyperplane separating classes cannot be always drawn neatly (optimal). Therefore a margin is introduced around the hyperplane between the classes to relax the constraint, allowing some instances to be mislabelled while allowing a better result in general. Consequently, a parameter as Complexity (C) is considered to identify the best distance for the margin. The objective is to find a suitable value of C that achieves a small classification error and at the same time creates a large margin around classes. The typical values of C are exponents of of 10 (e.g. 0.01, 0.1, 1.0, 10.0, 100.0) that will provide adequate classification performance divergence. A conventional way to identify a good value of C is employing a Grid Search approach. The Grid Search approach sequentially tries various values of C in a range, and chooses the one which results in the highest accuracy in recognition (Pfister and Robinson, 2011).

2.4.2 Facial Expression of Emotions

Facial expressions has been studied in behavioural science for more than hundred years (Darwin, 1998) and play an important role in the communication of human emotions (Ekman and Scherer, 1984; Izard, 2013). In a series of cross-cultural studies (Ekman, 1999), it was observed that humans are capable of perceiving the six basic emotions (2.3.1). In these studies, it was also found that these emotions are expressed universally via a stable configuration of facial movements.

The Facial Action Coding System (FACS) developed by Ekman and Friesen in 1978 (Ekman et al., 1978), is a comprehensive and psychometrically rigorous method that has been used in many studies (e.g Mehu and Scherer, 2015; Pantic and Rothkrantz, 2000) for analysing facial movements and inferring emotional states. FACS describes all visually observable movements of the facial muscles in terms of 44 measurement units that are called Action Units (AUs), for example AU12 is Lip Corner Puller. AUs differ from facial muscles; for example a combination of facial muscles could describe a single AU. Furthermore, AUs are objective measures of facial signals and therefore they can be mapped to high-level interpretation processes such as emotions (Zeng et al., 2009).

Previous research has shown the association between AUs and categorical and dimensional descriptors of emotions (Mehu and Scherer, 2015). For instance, the emotional category of joy, positive valence was observed to be associated with the increased activation of the Cheek Raiser (AU6) (Mehu and Scherer, 2015). Another example is the combination of Inner Brow Raiser (AU1) and Brow Lowerer (AU4) which often occurs in emotional sadness (Cohn, 2006).

FACS was originally designed for human observers to decompose a portrayed expression into the specific AUs that describe an expression. However, despite the large success of FACS by trained humans, it is a costly and time consuming process (Burr, 2006). Consequently, automating FACS would be a cost-effective alternative for facial expression analysis.

The first step in automatic recognition of facial expressions is acquisition of face (for example through static images or video) and then extracting the relevant information such as AUs. Developments of sensors with cameras to capture geometric features of the face (such as eyes or brows) have advanced the research in the understanding of facial behaviours. For example, the Microsoft Kinect sensor (Microsoft, 2015b), which is equipped with both RGB camera and depth data, permits tracking different feature points in 3D space. Recent studies such as (Alabbasi et al., 2015; Mao et al., 2015) have used Kinect to capture facial features for emotion recognition. Other studies such as (Bahreini et al., 2016) used a webcam to stream facial data for later analysis.

2.5 Experiences of Online Shopping

Emotions influence customers' cognition, such as the perception of effectiveness and informativeness (Mazaheri et al., 2012) in online environments. In particular, positive emotions most likely contribute to the simple and default way of information-processing, whereas negative emotions lead to the scrutinised analysis of information (Clore et al., 1994). Emotions can also cause changes on the current flow of interactions undertaken by the user. For instance, if the user browses a website to identify certain information, and when the slow Internet connection impedes this goal, the user's frustration might lead her/him to search for information in another resource (Stickel et al., 2009). Consequently, research has attempted to explore specifically what features of e-retail environments induce emotional responses. For instance, some studies have investigated the impact of interface design (e.g. Porat and Tractinsky, 2012; Ward and Marsden, 2003)) and atmospheric cues such as background colour (Cheng et al., 2009), text colour, and music (Ding and Lin, 2012) to understand and evoke customers' emotional responses, which in turn have an influence on their cognitive states, decision making and their appraisal of the e-retailer (Porat and Tractinsky, 2012).

To assess emotional responses on online environments, the S-O-R model of Mehrabian and Russel (Figure 2.3) has been frequently used (e.g. Ding and Lin, 2012; Floh and Madlberger, 2013; Mazaheri et al., 2012). As it was stated before, an environmental stimulus (S) incites user's internal emotional appraisal (O), which in turn leads to the user's reaction or response (R). Within a website, signs, symbols, artefacts, ambient conditions, space and function are reported as stimuli (Floh and Madlberger, 2013). Research has also observed that computer-related events are not essentially different from other stimuli, they cause physiological changes and in turn these changes are expressed as level of arousal and manifest clues for pleasure (Ward and Marsden, 2003). In addition to pleasure and arousal, dominance has been demonstrated as a reliable indicator for the controlling power over the environment in online context (Kahneman et al., 2003).

2.6 Conclusions

UX is a broad and intriguing research area in the field of HCI. Hassenzahl (2008) has defined UX as "momentary, primarily evaluative feelings (good-bad) while interacting with a product or service". From this definition three aspects can be inferred: First, UX as opposed to usability has extensively focused on the aspect of users' emotions. Second, these emotions are short-lived, and third they are triggered as the cognitive evaluation of events.

Despite the increasing interest in UX research, the commonly applied methods in UX studies have been mostly remained to the use of traditional techniques in psychology and usability. In other words there is a need for empirical research in order to investigate the integration of the more advanced approaches, especially with regard to measuring emotional experiences. Although there has been a number of contributions with the use of recognition techniques specifically through the use of facial expressions, however with lacking the theoretical background on how an overall experience can be inferred from momentary emotional expressions.

As a result, this thesis first seeks to bridge the existing gab between the methodology prevalently used in UX and state-of the-art techniques in information technology in general. Second, to exploit previous research in order to clarify how emotional experiences form during a course of interaction, and third, to provide empirical evidence to support the methodology proposed in this research.

2.7 Chapter Summary

In this chapter it was discussed that emotions are the essential components of human experience and UX as a subcategory of experience is created through technology usage. In this regard, positive experiences are the main goal of digital technology,

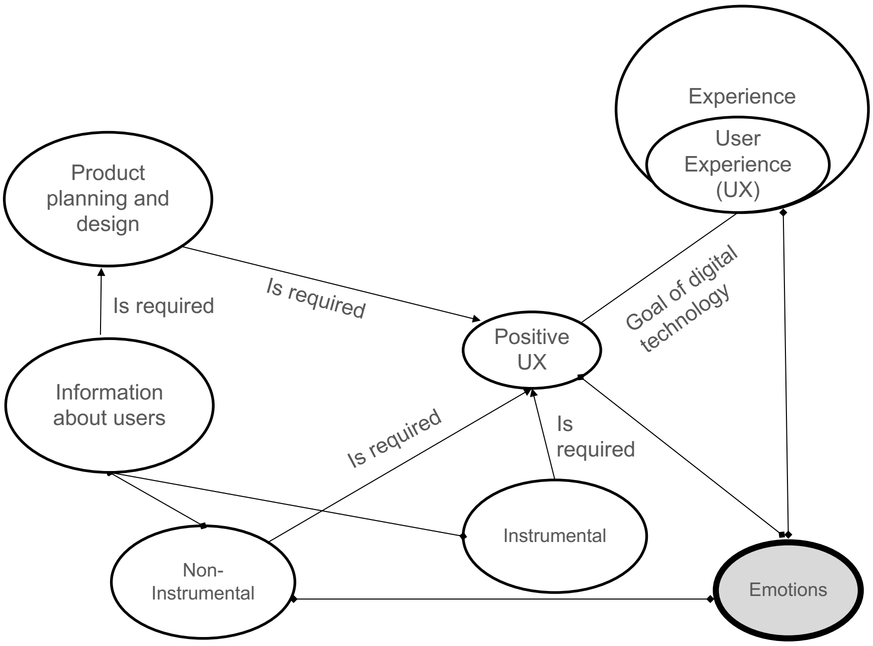


Figure 2.5: Overview of the forming relations between concept of UX, emotions and positive UX.

which can be resulted from fulfilling instrumental and in particular non-instrumental needs. Consequently, meeting users's needs requires product planning, gathering information about users and therefore conducting UX and usability evaluation sessions (Figure 2.5).

With respect to the approaches for evaluating UX, the methods can be divided into two camps: traditional and non-traditional methods. Traditional approaches have been originally borrowed from the usability techniques; such as interviews, user observations, focus groups and think aloud. While these approaches are usually employed retrospectively, they provide subjective reports of UX and also have the advantage of eliciting the overall perception of an experience. On the other hand, the non-traditional measurement methods include the use of sensors (such as camera, EEG sensors and eye-tracker) to capture both voluntary (e.g., facial expression) and involuntary bodily responses (e.g., brain waves) during an interaction. Therefore, a balanced use of both types of approaches could obtain a holistic view of UX.

Additionally, the think aloud protocol was explained. It was discussed that the original theoretical basis of think aloud protocol do not conform to the actual practice of that. To fill the gap between theory and practice, the speech communication approach can be used, as a relaxed way of the original think aloud. Last but not least, emotions were defined as evaluative responses towards internal and external events and as a consequence emotional cues manifested in vocal and facial expressions can be utilised for recognition by automated approaches.

Given the background of UX and its prevalent assessment methods (Section 2.2), next chapter describes the methodological approaches adopted in this thesis for measuring emotional experiences. The adopted methodologies have been derived from the affective computing and ML approaches.

Chapter 3

Recognising Vocal Emotions by Acoustic Analysis

This chapter describes the employed methodology for measuring the emotional aspect of UX by adopting an automatic approach in analysing vocal expressive emotions. As discussed in Chapter 2, the think aloud protocol as one of the extensively used methods in usability studies can be used for evaluating UX as well. To assess emotional expressions elicited during thinking aloud sessions, a practical and unobtrusive approach could be pursued by utilising techniques in machine learning (ML). In this regard, learning algorithms are trained based on previously annotated emotional corpora. Subsequently, emotional states of unknown data can be recognised by applying the generated classification models. This chapter begins with providing a summary of relevant methodologies for building a voice emotion recognition application. Next, the proposed approach in this work for analysing emotions will be explained.

3.1 Available Emotional Speech Corpora

Employing training data is a necessary prerequisite for detecting unknown emotions in an automatic recognition system (Ververidis and Kotropoulos, 2006). During the past two decades, several emotional corpora have been recorded to target different application purposes. A number of study reviews (such as El Ayadi et al., 2011; Ververidis and Kotropoulos, 2006; Wu et al., 2014) has surveyed the existing emotional corpora. Additionally, the Association for the Advancement of Affective Computing (AAAC) has listed a record of vocal, visual and multimodal emotional databases (Douglas-Cowie and members, 2004).

Research on the existing vocal emotional corpora suggests that although a variety of emotional databases has been developed by different research communities, most of these databases have not been available for public use (this issue has been reported by previous research as well (El Ayadi et al., 2011)). At the same time, many of the audio emotion datasets have been recorded as a part of a bigger audiovisual database. Consequently, from reviewing the surveys it is not clear whether the audio part of the database has been recorded and emotionally annotated separately from the visual part or not, and thus the corpus can be used for voice emotion recognition specifically (e.g. Wu et al., 2014). Hence, a more up-to-date review of the publicly available vocal emotional corpora could facilitate the related research in voice-based emotion analysis.

Reviewing the related literature and performing Google searches, eighteen emotional corpora were identified as publicly available datasets. The corpora and their characteristics have been summarised in Tables 3.1, 3.2 and 3.3. The following information is presented for each data collection:

1. Name of database
2. Accessibility (free or with licence fee)
3. Language of recording
4. Size (the number of participants and data samples; e.g. number of utterances)
5. Emotion elicitation method such as simulating emotions (by acting or posing), inducing emotions and the use of natural emotions
6. Emotion description (categorical labels, dimensional descriptors)
7. Modality: Audio or Audio-Visual
8. Publication year

From these tables it can be noticed that the available corpora are limited to a few number of datasets for each of the languages: English (7 corpora), German (5 corpora), French (2 corpora), Danish (1 corpus) and Chinese (1 corpus).

Among the listed corpora, AVIC (Schuller et al., 2007b) was in a particular interest for the data analysis in this thesis. The spoken language of AVIC database (Schuller et al., 2009a) is English and it contains natural dialogues. As the scenario, an experimenter and a participant interacted with each other, while the experimenter played the role of a product presenter and led the participant through a discussion on advertisement. The participant's role was to communicate with the presenter and express his/her interest in the topics. The labelling scheme was based on the perceived "Level of Interest" (LOI) on participants' utterances. Three LOIs were

Corpus	Access	Language	Size	Elicitation method	Emotions	Moda-lity	Year
Geneva Vocal Emotion Expression Stimulus Set (GVEESS)(Banse and Scherer, 1996)	Public and free	German	244 utterances, 12 actors (6 female, 6 male)	Simulated	Categorical labels; hot/cold anger, panic fear, anxiety, despair, sadness, elation, happiness, boredom, interest, shame, pride, disgust and contempt	A	1996
Danish Emotional Speech (DES)(Engberg et al., 1997)	Public with licence fee	Danish	419 utterances, 4 actors (2 female, 2 male)	Simulated	Categorical labels: anger happiness, neutral, sadness and surprise.	A	1997
Speech Under Simulated and Actual Stress (SUSAS)(Hansen, 1996)	Public with licence fee	English	16,000 utterances, 32 actors (13 female,19 male)	Simulated and Natural	Four stress styles (including amusement park roller-coaster and helicopter cockpit recordings)	A	1999
Berlin Emotional Speech Database (EMO-DB)(Burkhardt et al., 2005)	Public and free	German	535 utterances, 10 actors (5 female, 5 male)	Simulated	Categorical labels: anger boredom, disgust, fear, joy, neutral and sadness	A	2000
Belfast Sensitive Artificial Listener (SAL)(Douglas-Cowie et al., 2007)	Public and free	English	1692 utterances, 4 subjects (2 female, 2 male)	Induced (subjects talk to artificial listener with 4 different personalities)	Dimensional labels: valence arousal (annotating in a range from -1 to +1)	AV	2002
LDC Emotional Prosody Speech and Transcripts Corpus (Liberman, 2002)	Public and free	English	1050 utterances, 8 actors (5 female, 3 male)	Simulated	Categorial labels: disgust, panic, anxiety, hot anger, cold anger, despair, sadness, elation, happy, interest, boredom, shame, pride, contempt and neutral.	A	2002

Table 3.1: Characteristics of public emotional speech databases. A= Audio, V=Visual.

Corpus	Access	Language	Size	Elicitation method	Emotions	Moda-lity	Year
eNterface Database (Martin et al., 2006)	Public and free	English	1277 utterances, 42 speech-laughs 42 subjects (8 female, 34 male)	Induced	Categorical labels: anger, disgust, fear, happiness, sadness and surprise.	AV	2005
Electromagnetic Articulography Database (EMA) (Grimm et al., 2007)	Public and free	English	680 utterances, 3 subjects (2 female, 1 male)	Simulated	Categorical labels: happy, angry, sad, neutral, surprise and others.	A	2005
Geneva Multimodal Emotion Portrayals (GEMP) (Bänziger et al., 2006)	Public and free	French	1260 utterances, 10 actors (5 female, 5 male)	Simulated	Categorical labels: pride, joy, amusement, interest pleasure, relief, hot anger, panic, fear, despair, shame, irritation, anxiety, sadness, tenderness, contempt, surprise, admiration and disgust.	AV	2006
Interactive Emotional Dyadic Motion Capture (IEMOCAP) (Busso et al., 2008)	Public and free	English	10,039 utterances, 10 actors (5 female, 5 male)	Simulated and and Natural (actual) stress	Categorical labels: Anger happiness, sadness, frustration and neutrality. Dimensional labels: valence, arousal and dominance	AV	2007
Audiovisual Interest Corpus (AVIC) (Schuller et al., 2007b)	Public and free	English	3002 utterances, 21 subjects (10 female, 11 male)	Natural (human -to-human conversation)	Level of interest (LOI). LOI1: boredom, LOI2: neutral, LOI3: joyful.	AV	2007
Vera Am Mittag Database (VAM) (Grimm et al., 2008)	Public and free	German	1018 utterances, 47 speakers (36 female, 11 male)	Natural	Dimensional labels: valence, arousal, dominance	AV	2008
Airplane Behavior Corpus (ABS) (Schuller et al., 2007b)	Public and free	German	431 utterances, 4 speakers (2 female, 2 male)	Induced	Categorical labels: tired aggressive, intoxicated, nervous and neutral	AV	2009

Table 3.2: Characteristics of public emotional speech databases. A=Audio, V=Visual.

Corpus	Access	Language	Size	Elicitation method	Emotions	Modality	Year
The FAU Aibo Emotion Corpus (Batliner et al., 2008)	Public and free for scientific use	German	18,216 utterances, 51 children (30 female, 21 male)	Natural (Interaction with Sony's pet) robot Aibo)	Categorical labels: joyful surprised, motherese, neutral, bored, emphatic, helpless, irritated, reprimanding and angry.	A	2009
Surrey Audio-Visual Expressed Emotion (SAVEE) (Haq et al., 2009)	Public and free	English	480 utterances, 4 actors (4 male)	Simulated	Categorical labels: anger disgust, fear, happiness, sadness, surprise and neutral	AV	2009
RECOLA (Ringeval et al., 2013)	Public and free	French	46 subjects, (27 female, 19 male)	Natural	Dimensional labels: valence arousal (annotating in a range from -1 to +1). Social behavioural labels: agreement, dominance, engagement, performance and rapport.	A	2013
The MAHNOB Laughter Database (Petridis et al., 2013)	Public with licence fee	English	849 utterances, 67 speech-laughs 22 subjects (10 female, 12 male)	Induced and posed	Laughter	AV	2012
Subset of Audio-Visual Depressive Language Corpus (AViD-Corpus) (Valstar et al., 2013)	Public with licence fee	German, English	340 video-clips, 292 subjects	Natural (HCI tasks)	Dimensional labels: Continuous valence, arousal and dominance.	AV	2013
CHEAVD: a Chinese natural Emotional Audio-Visual Database (Li et al., 2017)	Public and free	Chinese	1160 utterances, 238 speakers (113 female, 125 male)	Natural (films and TV series)	Categorical labels: anger, joy, fear, frustration, sadness and surprise.	AV	2015

Table 3.3: Characteristics of public emotional speech databases. A=Audio, V=Visual.

defined: LOI0, indicated disinterest or when participant was tired of listening and talking about the topic, LOI1 showed neutrality or indifference, and LOI2 represented strong interest, where a participant was curious to discuss the topic and asked questions. In the context of HCI, recognising different levels of interest given to different aspects of interactive products/services can be very beneficial for UX researchers and practitioners, where positive and negative aspects of products/services could be identified accordingly. Results of emotion recognition can help identify the negative and positive elements of the product, thereby informing the redesign of the product or future development strategy for similar products.
Therefore, due to the mutual interest in designing this corpus and the objectives of this thesis, multiple requests were sent to get a hold of the AVIC corpus. However, no copy of the corpus could be obtained for conducting the studies of this PhD research.

Different objectives and application purposes underlying the development of a corpus result in different characteristics of the corpus (El Ayadi et al., 2011). For instance, the goal of the recognition task influences the design of the corpus in terms of the type and number of the target emotions (e.g. recognition of frustration or stress in call centre applications (Ma et al., 2006)). However, despite different characteristics, the literature shows that four general criteria are usually considered in the evaluation and selection of a corpus:

1. **Emotion elicitation methods**: Elicitation methods for collecting the content of emotional databases could be classified into three major categories:

   Simulated: The first category includes the approaches that simulate speech with different emotional states. In this regard, professional or nonprofessional actors are asked to express (pose/act) each emotion deliberately. As it can be seen from the review figures above, most of the existing databases have collected simulated emotions. Simulating emotions by professionals is described as the most reliable method to collect emotionally coloured data (Ververidis and Kotropoulos, 2006). On the other hand, the problem with simulated databases is that actors express the emotions in an exaggerated way, reducing the degree of naturalness of the database (El Ayadi et al., 2011). To overcome the problem of unnaturalness, some databases such as the Berlin Emotional Speech Database (Burkhardt et al., 2005) made use of nonprofessional performers to act the emotional states. Another issue related to acted emotions is that to obtain speech data, scripted texts are used by performers to read. However, according to some research findings (Williams and Stevens, 1972), the speech produced by reading scripts differs from spontaneous speech.

Induced: The second category includes induced emotions, which are collected within speech in artificial situations by the use of clips, music, scenarios, etc. For example, eNTERFACE'05 EMOTION Database (Martin et al., 2006) was collected during the eNTERFACE'05 workshop. In this workshop participants were asked to listen to six different stories, each inducing a specific emotion. Induced emotions are neither natural nor simulated (Ververidis and Kotropoulos, 2006). Thus, induced emotions are more realistic than simulated emotions. However, induced emotions are still confined to the elicitation scenario.

Natural: The third category is natural emotions elicited spontaneously and in real-life situations. Examples of natural emotional databases are the recording of call centre interactions (Bennett and Rudnicky, 2002) and TV-recordings, The Vera-Am-Mittag (VAM) database (Grimm et al., 2008). Although such recordings contain natural emotions, ethical or privacy issues are associated with the use of them. Additionally, manual labelling of spontaneously expressed emotions is very expensive, error prone and time consuming (Zeng et al., 2009).

In some literature (e.g. Schuller et al., 2011a), simulated emotions are referred to as prompted whereas induced and natural emotions are referred to as non-prompted emotions. To produce non-prompted emotions, a scenario for recording can be through human-human or human-machine interaction. In the human-machine interaction, a human operator can play the role of machine. In this case the scenario is called Wizard-of-OZ (WoZ). Examples of databases recorded through WoZ scenarios are Belfast Sensitive Artificial Listener (SAL) (Douglas-Cowie et al., 2007) and FAU Aibo Emotion Corpus (Batliner et al., 2008).

2. **Emotion labelling**: There are two major methods to encode emotions: Categorical or discrete labelling such as joy, anger, and surprise, and dimensional labelling along different dimensions such as valence, arousal and dominance. Apart from these two descriptors, some studies have used behavioural labels (e.g. interest) and social labels (e.g. agreement and performance Ringeval et al., 2013). From the literature it can be noted that categorical labelling dominates the majority of the databases. In this case five or six basic emotion labels - anger, joy, disgust, sadness, fear and neutral - are used for annotating the content. On the other hand, a fewer number of studies have used dimensional labelling such as in (Douglas-Cowie et al., 2007).

3. **Emotional units**: Defining appropriate speech episodes as "units of analysis" is a precondition to analyse the emotional content. In acted data, the emotional units are usually discrete, constant and short episodes. In conversations/interactions, the units are often dealt as 'turns', which start when a speaker starts speaking and end when she/he stops speaking. In more naturalistic data, emotional contents may not be discrete and sometimes are very long. In such cases, other measures must be considered.

   One approach is employing objective measures to identify the meaningful units with varying lengths; for example units with predefined length (e.g. 5 seconds length) of duration (ibid.), although there is no clear consensus concerning the best length of duration for emotion analysis (Ringeval et al., 2015). Using pauses for longer than a specific threshold (such as 0.5 or 1 second Schuller et al., 2011a) is another measure to define unit of analysis (Rahman and Busso, 2012). However, these objective measures can result in segments that do not necessarily comply with emotional units (Schuller et al., 2011a).

   The other commonly used approaches include identifying units with varying lengths and time intervals such as words or utterances. An utterance has been described as a semantically and linguistically well-defined chunk with no change of emotion within it (Vogt et al., 2008). An utterance could consist of a single word (i.e. "Yup" and "Okay") to multiple sentences.

4. **Size**: The size of the database both in terms of number of participants taking part and number of utterances recorded plays an important role in characterising qualities of a database. For example, the properties such as generalisability and scalability of a database are associated with the number of participants involved in the creation of a database (Koolagudi and Rao, 2012). Another consideration is the number of utterances for each emotion. Some database developers prefer the same number of utterances for each emotion in order to have a clearer evaluation of how a database performs on classifying emotions (such as the EMA database Grimm et al., 2007). On the other hand, some other developers argue for an unbalanced distribution of emotions in a database to better reflect the occurrence of emotions in daily life situations (Hansen, 1996). For instance, since the neutral emotion is the most frequent emotional state in everyday life, therefore the number of utterances with neutral emotion should have the largest number in the database (El Ayadi et al., 2011).

3.2 Implementation Methodology of Voice Emotion Analysis

3.2.1 Selected Emotional Corpora

To set out the general implementation methodology of voice emotional analysis, four corpora were chosen: One German, EMO-DB, and three English, SAVEE, EMA and IEMOCAP. These corpora were among the most popular and publicly free available emotional speech datasets. The rationales behind this selection attain to the publicly availability of the corpus, language and the applied method for emotion labelling. An overview of the properties of the selected corpora can be found in Tables 3.1, 3.2 and 3.3. Additionally, a more detailed description of each of the selected corpora is provided as follows:

- **Berlin Emotional Speech (EMO-DB)**: EMO-DB (Burkhardt et al., 2005) is one of the most well-known databases that has been utilised frequently in the field of emotion classification. This database contains acted emotions in seven categories: anger, happiness, sadness, boredom, disgust, fear and neutral mode. In total, there are 535 utterances for all emotional labels in the database. The content of this database is predefined and consists of ten German sentences that are semantically neutral (An example of a sentence in this corpus is: "Der Lappen liegt auf dem Eisschrank", which is translated to English as: "The tablecloth is lying on the fridge"). The mean accuracy of the perceived emotions obtained through human evaluation is about 84% for this database. The EMO-DB database has been used as a benchmark in a number of emotion related studies and its performance accuracy is reported frequently. Thereby, I also uphold this selection in order to draw a comparable conclusion in the performance accuracy of different corpora.

- **Surrey Audio-Visual Expressed Emotion (SAVEE)**: The SAVEE corpus (Haq et al., 2009) is a public, audiovisual emotion database. This database consists of seven acted emotions: anger, happiness, disgust, fear, sadness, surprise and neutral. Four male English actors took part in the recordings of this database and in total 480 utterances were collected (120 utterances per actor). To create the emotional contents, phonetically balanced and predefined sentences were chosen for each emotion. The average accuracy obtained by human evaluation for this corpus was 65.5%.

- **Electromagnetic Articulography (EMA)**: The EMA corpus contains 680 utterances of four different emotions - anger, happiness, sadness and neutrality, which were simulated by three native American English speakers, one male and two female (Lee et al., 2005). The EMA data was perceptually evaluated and the average accuracy of human evaluation was 81.9% (Grimm et al., 2007). In this database in addition to the acoustic signals, articulatory information was collected as well. However, for this research only the acoustic data is utilised.

- **IEMOCAP**: IEMOCAP is an audio-visual corpus generated by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (Busso et al., 2008). This database includes approximately 12 hours of dyadic interactions from ten trained actors (five male, five female). In this database, actors performed scripted and spontaneous emotional scenarios designed to evoke specific types of emotions, namely: happiness, sadness, anger, frustration and neutral state. These emotion labels were chosen in respect to the most common ones in the literature (Picard and Picard, 1997). However, during the evaluation of this database by human evaluators, the emotional categories were expanded to include excitement, disgust, fear and surprise as well. In addition to the discrete emotional labels, the content of this database was annotated with dimensional descriptors arousal [1 calm - 5 excited], valence [1 negative - 5 positive] and dominance [1 weak - 5 strong] for each utterance. Human evaluators assessed each utterance in terms of emotional categories and the average accuracy was 74.6%.

  Note: The size of IEMOCAP is approximately 10,039 utterances. However, the number of utterances has not been evenly distributed among the emotional states. For example, for the emotion labels fear, surprise and disgust, each comprises less than 1% of the entire database (less than 100 samples). For this thesis, those labels (such as disgust and surprise) were removed from the rest of the analysis. Hence, the final six emotional labels chosen were: Happiness, sadness, anger, frustration, excitement and neutral state. Moreover, using the entire database for training the ML algorithms required very long computational time (given the large number of feature sets and the complexity of SVM algorithms). Therefore, a random and relatively balanced subset of this database, including about 2000 utterances, was considered.

Despite the public availability and English language of Belfast Sensitive Artificial Listener (SAL) corpus (Douglas-Cowie et al., 2007), this corpus was not evaluated due to the reason that its content has only been labelled with respect to valence and arousal (in a continuous scale from -1 to +1). Not containing information about the third dimension, dominance, made this corpus unfitted for this work.

AViD (Valstar et al., 2013) was another interesting corpus that was found relevant. A subset of the AViD corpus includes recording of participants performing HCI

Feature Group	Features in Group
Raw Signal	Zero-crossing-rate
Signal energy	Root mean-square and logarithmic
Pitch	Fundamental frequency F0 in Hz
Voice quality	Probability of voicing
Spectral	Energy in bands 0-250Hz, 0-650Hz, 250- 650Hz, 1-4kHz
Mel-spectrum	Band 1-26
Cepstral	MFCC 0-12

Table 3.4: Examples of the Low Level Descriptors (LLD) of acoustic features implemented in openSMILE (Schuller et al., 2009b).

Functions
Range (max.-min.)
Arithmetic, quadratic, and geometric mean
Std. deviation, variance, kurtosis, skewness
Zero-crossing and mean-crossing rate

Table 3.5: Examples of statistical functionals implemented in openSMILE (Schuller et al., 2009b).

related tasks with a webcam and a microphone. In this corpus, participants were asked to speak out loud while solving a task similar to this research work. However, the German spoken language of this subset and a licence fee, also made this database unsuitable for evaluation and use in this work.

3.2.2 Feature Extraction

To extract acoustic features, the openSMILE (Speech and Music Interpretation by Large-Space Extraction) toolkit (Eyben et al., 2010) was used. OpenSMILE is an open-source signal processing and feature extraction toolkit that can generate more than 5000 acoustic features from an input audio in real time. The SMILE feature extractor is capable of extracting a set of low level acoustic features (such as pitch and energy) and applying statistical functionals (such as mean and variance) and transformation functionals (such as the derivatives) to those features. Table 3.4 and Table 3.5 respectively list examples of the acoustic features and applied functionals used in openSMILE.

OpenSMILE comes with a number of predefined feature sets (such as the sets for the INTERSPEECH Emotion Challenge). Depending on the task a suitable feature set can be configured for feature extraction.

As explained in Section 2.4.1, the standard and extended sets of EC contain 384 and 6552 features respectively. While various studies have used the standard EC for emotion recognition (e.g. Rahman and Busso, 2012; Rosenberg, 2012; Zhang et al., 2016), in the related literature, the extended set of EC has received less attention for evaluation (e.g. this set was tested in Pfister and Robinson, 2011).
To determine which set outperforms in terms of emotion recognition accuracy, I evaluated both standard and extended sets for feature extraction. Additionally, the eGEMAPS with 88 parameters, the most recent implemented baseline feature set was also chosen for evaluation. The eGEMAPS has been used in a number of recent studies for feature extraction and emotion recognition (such as Ringeval et al., 2016; Solera-Urena et al., 2017).

3.2.3 Classification and Evaluation

Incorporating four emotional corpora and three feature sets (4x3), 12 experiments were carried out: The emotional corpora EMO-DB, SAVEE, EMA and IEMOCAP were used to train the classification algorithms. The acoustic features of the corpora were extracted using openSMILE and according to the standard and extended EC, and the eGEMAPS sets.

For classification of emotions, Support-Vector Machines (SVM) with a linear kernel function were chosen. SVM classifiers have shown higher effectiveness than the other machine learning algorithms when a large feature set is being used (Shami and Verhelst, 2007). The SVMs were trained by the Sequential Minimal Optimisation (SMO) algorithm (Platt, 1998). For the complexity parameter C, a range of values was tested, where $C \in \{1, 0.5, 0.1, 0.05, 0.01\}$.
To evaluate the performance of a classifier, the general approach is to split data into training and testing parts. In this regard, data is split into N (e.g. N=10) number of folds, thus training on N-1 folds, and testing on the remaining data. This process is repeated N times, changing the tested partition in each iteration. This approach is called *N-fold cross validation*. When the distribution among classes is kept throughout the data partitioning, the approach is known as stratified cross validation. All the experiments in this research were carried out in a stratified 10-fold (N=10) cross-validation paradigm. Ten is the most common value for N used in the literature (Shami and Verhelst, 2007), therefore it was also adopted in these experiments.
Overall, the standard accuracy is one of the frequently reported measures for performance. Accuracy is the total number of correctly classified samples divided by the total number of samples in the testing set. Accuracy can also be referred to as

the weighted average of class-wise recall rates (WAR). However, the WAR cannot reflect an appropriate measure of performance, where there is the imbalance problem in the number of instances among the classes. An alternative way is obtaining the recall per class and reporting the unweighted mean among the classes (UAR). Another commonly reported metric is the precision rate, which is the fraction of the correctly classified samples divided by the total number of recognised samples from that class. Finally the F1 measure, which is the harmonic mean of of recall and precision. Recall (sensitivity) and precision (specificity) can be expressed by the terms true positives (TP), false positives (FP), and false negatives (FN) as well. TP is the number of correctly identified samples, FP is the number of incorrectly identified samples and FN is the number of incorrectly rejected samples. In this regard, recall is the number of true positives over the number of true positives plus the number of false negatives 3.1, while precision is the number of true positives over the number of true positives plus the number of false positives 3.2.

$$Recall = \frac{TP}{TP + FN} \tag{3.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

A representation of all the TP, FP, FN and TN (true negative or correctly rejected samples) metrics is a table called Confusion Matrix. The WEKA toolkit generates confusion matrix as an output of the classifier evaluation. An example of a confusion matrix is given in Table 3.8, which will be explained in Section 3.3.

Table 3.6 shows the classification results obtained for each corpus with respect to discrete emotion categories. The UAR was used as the measure of performance due to the imbalanced number of instances in the different classes of the EMO-DB corpus. Basically, the UAR and WAR reflects the same measure, where different classes contain a balanced number of instances.

According to the literature, the number of acoustic features has an important role in the classification performance (Schuller et al., 2009b). Experiments carried out by Schuller and colleagues resulted in better recognition accuracies using sets with larger numbers of acoustic features (Schuller et al., 2009b). Given the means of the obtained accuracies using the three feature sets (Table 3.6), using the extended EC demonstrated mostly better recognition performances as compared to the other sets. Consequently, this set was chosen for feature extraction for the rest of the analysis in this work.

Corpus	Standard EC	Extended EC	eGEMAPS
EMO-DB (7 emotions)	81.2	88.6	80.0
EMA (4 emotions)	99.4	97.1	92.6
SAVEE (7 emotions)	82.3	86.4	69.2
IEMOCAP (6 emotions)	52.4	58.8	52.1
Mean	83.6	87.7	73.5
SD	19.5	16.6	17.1

Table 3.6: Unweighted average of class-wise recall (UAR) rates obtained for categorical emotion recognition by the use of SVM. 10-fold Stratified Cross-Validation. Using standard and extended INTERSPEECH 2009 Emotion Challenge (EC), and eGEMAPS feature sets, with respectively 384, 6552 and 88 acoustic parameters.

In machine learning it is a well-known fact that the more heterogeneous the training corpus is, the lower accuracy should be expected from the classification (Shami and Verhelst, 2007). In addition, classification performance is dependent on the type of database (for example whether the database is acted or spontaneous). The IEMOCAP database consists of spontaneous content as well as a higher number of speakers in its recording in comparison to the other corpora. Hence, the lower classification results on IEMOCAP compared to the other databases could be explained based on the given reasons. In other words, the corpora with acted data yielded a higher performance for recognition, which has been observed in previous work too (Schuller et al., 2009b).

To summarise, the overall implementation methodology of analysing emotions by vocal signals was explained. The process involved training the classification models based on the patterns found for different emotions in acoustic features. Particularly in this study, three publicly free available English emotional corpora and one German corpus (as the benchmark) were evaluated for emotional recognition. EMO-DB has been used in a number of related studies as the benchmark (such as Eyben et al., 2016). It should be noted that the recognition results obtained for EMO-DB are comparable with previous work (such as Eyben et al., 2009). Therefore, the comparable recognition results obtained for this database with previous research could verify the validity of the procedure followed in this study.

Adopting the described overall methodology, details of two separate experiments will be discussed in the following sections. The first experiment (Section 3.3) was carried out using one corpus, IEMOCAP. This corpus consists spontaneous data and has a large number of utterances. Those two factors made an optimal option for emotional analysis of data elicited in natural settings such as the verbalised data in thinking aloud. Subsequently, a second experiment (Section 3.4) was undertaken to investigate a combination of methods for improving the accuracy performance from the first experiment.

3.3 Classification Models based on IEMOCAP Corpus

The IEMOCAP corpus was chosen for training the classification models due to a number of reasons: First, the corpus has been annotated with discrete and dimensional (valence/pleasure, arousal and dominance) descriptors. Second, the content of this database consists scripted scenarios as well as improvised dialogs in spontaneous situations. Additionally, in the design of the IEMOCAP corpus, ten different actors were used. Despite the fact that ten is still a small number of speakers, it was expected that with having a higher number of actors as compared to the other corpora (such as EMA and SAVEE), the plausibility of analysing data for emotion recognition would be increased considering the inter-personal differences among diverse users.

The dimensional annotation of the content of IEMOCAP could provide the skeleton for generating models for emotion recognition based on the PAD model proposed by Mehrabian (1974). The dimensional annotation in this corpus was performed using the SAM scales ranging from 1 (lowest) to 5 (highest) for each of the three dimensions. The prominent superiority of using the PAD model relates to the potential of this model for describing complex and mixed emotions in the three nearly independent continuous dimensions: Pleasure/Valence (P), Arousal (A), and Dominance (D).

Following the procedure of emotion recognition, three classifiers were generated for classifying the three dimensions based on the PAD model. In this regard, the openSMILE toolkit and INTERSPEECH 2009 EC feature set were employed for feature extraction. Next, the classification models were trained by SVM algorithms with SMO linear function, which have been implemented in the WEKA software. Moreover, in order to optimise the recognition results of SVM classifiers, in a 10-fold stratified cross-validation paradigm, a range of values for C (the complexity parameter) was tested through a grid search mechanism. As it was described before, this is a required step to identify a good value of the margin around the lines between classes.

Additionally, using the IEMOCAP corpus, a separate classifier was created for recognition based on discrete emotions (Table 3.7). As it was explained before, the chosen

Discrete Category	Dimensions		
	Valence	Arousal	Dominance
60.1	62.9	69.1	60.3

Table 3.7: Unweighted average of class-wise recall (UAR) obtained for the discrete and dimensional emotions -valence, arousal and dominance- using SVMs in a 10-fold Stratified Cross-Validation.

	Classified				
Classes	a	b	c	d	e
Happiness = a	237	24	27	41	32
Anger = b	37	171	9	61	18
Sadness = c	15	5	239	20	11
Frustration = d	60	59	45	141	35
Neutral = e	40	20	22	39	157

Table 3.8: Confusion matrix of the classification results on the discrete emotions of the IEMOCAP corpus. Grey diagonal cells show the number of correctly classified instances per class.

discrete categories initially were: Happiness, sadness, anger, frustration, excitement and neutral state. However, due to the similarity between the happiness and excitement emotion categories, the instances of these classes were merged into one class (a significant overlap had also been observed within human annotators' evaluations in annotating the IEMOCAP corpus (Busso et al., 2008)).

Previous research Rahman and Busso (2012) had reported accuracy of 69.8% for a binary classification problem (emotional versus neutral) on the IEMOCAP corpus. Given the five classification problem for each of the discrete emotions, and dimensions valence, arousal and dominance, these results confirm the challenging task of emotion recognition on this database.
Table 3.8 shows the confusion matrix of the classification results on the discrete emotions of the IEMOCAP corpus. The results presented in this table reveal that there are more confusions in identifying samples between certain classes. For example, the instances of the anger class were mainly confused with the happiness (FN= 37) and frustration (FN=61) samples. The same confusions can be observed for the frustration class of which the samples were falsely classified as happiness (FN= 60) and anger (FN= 59). Confusion between the frustration and anger classes were also observed between human evaluators when annotating this corpus.

Regardless of the recognition performance, a caveat of the reported classifications is that the training and testing data came from one corpus. Therefore, the actual performance of the classifiers on real unknown data cannot be reflected from the given recognition results. These shortcomings will be addressed in the second sets of experiments.

3.4 Classification Models based on Cross-Corpora and Emotion Clustering Approach

In many emotion recognition systems, the evaluation performance is usually based on utilising one corpus for training and testing (Schuller et al., 2010b). However, this simplification cannot be generalised in real-life applications, where the recording settings (such as speakers group and microphone positions), emotion elicitation methods, annotation schemes and spoken language could largely vary. To tackle this issue, one solution is performing evaluation based on a cross-corpora approach, where classification models are trained on one set and tested on a completely different set. Consequently, it is expected to obtain a more realistic view of emotion recognition on unseen data by adopting the cross-corpus approach.

One specific problem associated with cross-corpus emotion recognition is the different emotion labels used for annotation. In this regard, each emotional corpus is usually recorded to address a specific task (for example the SUSAS corpus was created for recognition on the stress level (Hansen, 1996)). Therefore each corpus only contains a specific subset of emotions assigned to the spoken utterances. To address the problem of variability of emotion categories among corpora, one approach is to limit the extent of the evaluation to those emotion labels that are present in all data sets. However, this approach leads to the exclusion of specific emotion labels.

On the other hand, clustering emotions into general categories can cope with the variation of emotional labels in different corpora. For instance, the dimensional description of emotions can be used to map (cluster) discrete emotion labels to the general binary categories arousal (low/high), valence (negative/positive) and dominance (low/high). Clustering emotions into dimensions has the advantage of deriving a comparable insight of the accuracies obtained among corpora with different annotation schemes. Furthermore, by this approach, a higher number of different emotion labels can be incorporated, hence no need for the exclusion of emotion labels (Valstar et al., 2013). Finally, having binary clustering reduces human effort for labelling ambiguous emotions.

Despite the use of the binary clustering approach in a number of studies (e.g. Marchi et al., 2015a; Schuller et al., 2013, 2009b), there is a lack of justification for mapping some emotion labels such as the neutral state to the high-level emotion dimensions (e.g. high vs. low or negative vs. positive). In some emotion related works, such

	Valence	
Arousal	Positive	Negative
	Elation (joy)	**Hot anger/rage**
High	Amusement	Panic fear
	Pride	Despair
	Pleasure	**Cold anger/irritation**
Low	Relief	Anxiety/worry
	Interest	Sadness/depression

Table 3.9: Mapping of the emotion categories in GEMEP to binary valence, arousal and dominance. Emotions in high dominance are in bold (Adopted from (Mehu and Scherer, 2015)).

as the GEMEP corpus (Bänziger and Scherer, 2007), the neutral state has not been included as a label in the dataset.

In the GEMEP corpus, Scherer and colleagues provided a mapping scheme for the emotion categories to binary valence, arousal and dominance (Table 3.9). As it can be seen from the table, neutral state has not been considered in the binary mappings.

On the other hand, incorporating neutral state is necessary for applications, where this state occurs frequently (e.g. call centres Vaudable et al., 2010). Subsequently, Schuller and colleagues categorised neutral state as low arousal (or class 0 of arousal) and as positive valence (or class 1 of valence) for binary emotion classification (Schuller et al., 2009b).

Examples of mapping discrete emotions to binary valence and arousal can be seen in Table 3.10 and Table 3.11 for the four databases EMO-DB, EMA, SAVEE and IEMOCAP. The mapping of the emotion labels of these corpora have been adopted from Schuller (2009b).

Although there is no theoretical background for the current mappings of neutral state to low arousal and positive valence, related research supports the reliability of these mappings to some extent. For example, perviously it was observed that the mean value of the energy feature of neutral state is similar to sadness (low arousal) has a lower intensity compared to the energy feature of happiness and anger (high arousal). Likewise, pitch-related features (such as the pitch level and the pitch range) of happiness and anger are relatively higher than the corresponding features of sadness and neutral state (Wu et al., 2013).

While the literature has provided some direction for mapping neutral state to the binary arousal and valence, there is no previous research for mapping neutral state to binary dominance. Kouklia and colleagues (2013) showed that the perception of

Corpus	Negative (0)	Positive (1)
EMO-DB	Sadness, Anger, Fear Disgust, Boredom	Happiness, Neutral
EMA	Sadness, Anger	Happiness, Neutral
SAVEE	Sadness, Anger Fear, Disgust	Happiness, Neutral Surprise
IEMOCAP	Sadness, Anger Frustration	Happiness, Neutral Excited

Table 3.10: Mapping of discrete emotions to binary valence.

Corpus	Low (0)	High (1)
EMO-DB	Sadness, Neutral Disgust, Boredom	Happiness, Anger Fear
EMA	Sadness, Neutral	Happiness, Anger
SAVEE	Sadness, Neutral Disgust	Happiness, Anger Fear, Surprise
IEMOCAP	Sadness, Neutral	Anger, Happiness, Frustration, Excited

Table 3.11: Mapping of discrete emotions to binary arousal.

neutral state with respect to dominance can change depending on the contextual conditions of the recorded utterances. For example, in their study it was observed that human evaluators rated neutral state as a high dominance state, where utterances were recorded within a conversational scenario, and as a low dominance state, where utterances were recorded within a non-contextualised condition (such as acted data when recorded through reading sentences with specific emotions).

As a consequence of the lack of research on clustering the dominance dimension, this research will undertake a different approach for analysing this dimension from vocal cues. In the following subsections, first arousal and valence will be addressed, next the adopted approach to analyse dominance will be described (Section 3.4.2).

3.4.1 Binary Approach on Arousal and Valence Classifications

Using the four emotional corpora and binary clustering of emotion labels, the profile of the number of instances mapped to each class of valence (negative/positive) and arousal (low/high) can be found in Table 3.12.

Following the standard procedure of classification, SVMs were trained to classify the instances of each of the emotional corpora to low and high arousal; and to negative

Corpus	Arousal		Valence	
	Low (0)	High (1)	Negative (0)	Positive (1)
EMO-DB	267	268	385	150
EMA	340	340	340	340
SAVEE	240	240	240	240
IEMOCAP	737	790	655	677

Table 3.12: Number of instances in each class of binary arousal (low/high) and binary valence (negative/ positive).

Corpus	Arousal			Valence		
	Low (0)	High (1)	Accuracy	Negative (0)	Positive (1)	Accuracy
EMO-DB	0.97	0.96	96.3	0.91	0.82	88.4
EMA	1.00	0.99	99.2	0.95	0.94	94.9
SAVEE	0.95	0.93	93.9	0.73	0.78	76.0
IEMOCAP	0.67	0.66	66.7	0.63	0.64	63.2

Table 3.13: Class-wise recall obtained for the two dimensions - arousal and valence using SVMs. 10-fold Stratified Cross-Validation.

and positive valence. As it was explained before, 10-fold stratified cross validation technique was employed. Results of the classification can be seen on Table 3.13. Results show the same trend of recognition accuracy among corpora: Acted corpora resulted in better performances compared to the IEMOCAP corpus containing spontaneous and acted data.

Tables 3.14 and 3.15 summarise the results of performing cross-corpora evaluations based on the two dimensions arousal and valence. Similar to the results of previous studies (e.g Shami and Verhelst, 2007), classification performance dropped when training and testing sets were chosen from different corpora.

In addition to the UAR, Tables 3.14 and 3.15 present the geometric means (G-Mean) of the recalls of the two classes 0 and 1 obtained from the different cross-corpus evaluations. The G-Mean represents the mean value of different data by taking into account the possible skewness between the data. For instance, in a classification problem, a classifier could perform well in recognising the instances of a particular class, however poor on other classes. While in this case the arithmetic mean fails to reflect the unevenness that exists between the recall rates of the different classes, the G-Mean is sensitive to such skewness.

Furthermore, given the G-Means of the recall rates on arousal and valence (Tables 3.14 and 3.15), it seems that the classifiers trained from the EMO-DB corpus have the lowest classification performances when applied to the other corpora. One implication is that classifiers trained by German speech data cannot generalise well

Training	Testing	UAR	G-Mean
EMA	IEMOCAP	0.64	0.59
	SAVEE	0.52	0.52
	EMO-DB	0.48	0.18
	Average	0.55	0.43
IEMOCAP	EMA	0.61	0.59
	SAVEE	0.51	0.51
	EMO-DB	0.51	0.44
	Average	0.54	**0.51**
SAVEE	IEMOCAP	0.63	0.62
	EMA	0.57	0.42
	EMO-DB	0.51	0.15
	Average	0.57	0.40
EMO-DB	IEMOCAP	0.13	0.51
	EMA	0.52	0.18
	SAVEE	0.52	0.21
	Average	0.39	**0.30**

Table 3.14: Cross-corpus evaluation results on arousal based on unweighted average of class-wise recall (UAR) and geometric mean (G-Mean). The highest and lowest G-Means are shown in bold.

Training	Testing	UAR	G-Mean
EMA	IEMOCAP	0.51	0.43
	SAVEE	0.51	0.51
	EMO-DB	0.53	0.51
	Average	0.52	0.48
IEMOCAP	EMA	0.56	0.54
	SAVEE	0.52	0.52
	EMO-DB	0.44	0.47
	Average	0.51	**0.51**
SAVEE	IEMOCAP	0.51	0.16
	EMA	0.50	0.10
	EMO-DB	0.51	0.27
	Average	0.50	0.17
EMO-DB	IEMOCAP	0.50	0.00
	EMA	0.50	0.00
	SAVEE	0.50	0.54
	Average	0.50	**0.18**

Table 3.15: Cross-corpus evaluation results on valence based on unweighted average of class-wise recall (UAR) and geometric mean (G-Mean). The highest and lowest G-Means are shown in bold.

for recognition on English data. In other words, these results could indicate the impact of language on the classification results, when the language of the training set is different from the language of the testing set. Research (such as Abdelwahab and Busso, 2015; Schuller et al., 2010b) has previously suggested language, the conditions of the recording settings and speaker variations can cause drop in the recognition performance.
In addition, as can be seen from Table 3.15, the G-Means of the recall rates on IEMOCAP and EMA are 0.0 when the classifiers are trained based on the EMO-DB corpus. This is due to the 0.0 recall rate on recognition on the class low (0) of valence. In other words, the classifiers trained on the EMO-DB could not recognise the class low (0) of valence when they applied on the IEMOCAP and EMA corpora. While, the UAR rates failed to show the biased recognition between the two classes, the G-Mean rates reflected such biases.

Overall, a better recognition performance was observed for arousal in comparison to valence. In this regard, cross-corpus evaluations within English corpora showed above the chance level of recognition for arousal.

An Adaptive Approach by Concatenating Neutral References

As it was shown previously, having different training and testing conditions (such as differences in recording, speaker variation, etc.) resulted in dropping the accuracy of emotion recognition. This could be explained by the fact that same emotional labels in different databases occupy different regions of the feature space, which in turn leads to little generalisation across corpora (Shami and Verhelst, 2007).
To reduce the effect of these differences, in this section an approach is proposed for a model adaptation where the classifier is trained based on small subset of labelled data extracted from a target set (testing set) in addition to the original training set. In this regard, a small number of neutral labelled references from a target set is taken out and concatenated to the training set (see Figure 3.1). Consequently, it is expected that the classification error over the training examples is decreased. Note that concatenating neutral labels to a training domain relies on this assumption that these references are easily obtainable from individuals, especially in the context of recognising elicited emotions during a usability test session.

The use of adaptive algorithms has been proposed in previous work as well, where SVM classifiers were transformed to a new SVM classifier based on a small number of labelled data (Yang et al., 2007). Wu and colleagues proposed incorporating prior knowledge of some testing data, in which the data has a rough label and a corresponding confidence value, to generate SVMs (Wu and Srihari, 2004). In this

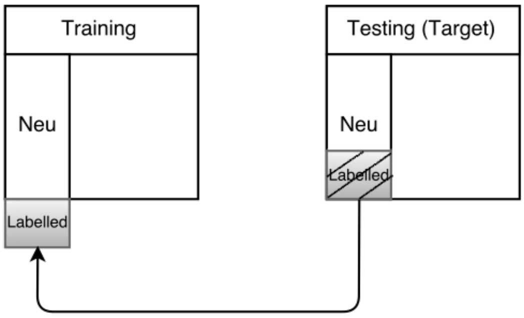


Figure 3.1: A schematic view of the proposed adaptive approach. A small number of neutral (Neu) samples from the target set are removed and concatenated to the training set.

regard, the confidence value of the labels influences the position of the margin for an SVM. Finally, the active learning approach in machine learning involves the process of adding new labelled data to the training set and updating the parameters of the learning algorithms for classification.

To test the performance of the proposed approach, a small number of neutral samples (e.g. 10% of all the neutral samples in a target corpus) are drawn out randomly and added to a training corpus for model creation. Among the other corpora, IEMOCAP was selected as the training dataset due to three reasons: First it contains spontaneous data, second, IEMOCAP has remarkably a higher number of instances compared to the other corpora EMA, SAVEE and EMO-DB. Finally, the number of actors in IEMCAP was more than the other corpora, resulting in a higher degree of generalisability. Consequently, EMA, SAVEE and EMO-DB were used as the target domains.

In this regard, from each of the target domains EMA, SAVEE and EMO-DB, a small number of neutral samples was extracted and added to the training set, IEMOCAP. Consequently, three corpora pairs were considered for evaluating the performance of the classification based on this adaptive approach: IEMOCAP (training) labelled-EMA (target), IEMOCAP (training) - SAVEE (target) and IEMOCAP (training) - EMO-DB (target).

With respect to the extraction of the neutral references, the adaptation training set was created at random from all the subjects in the target set. The size of the adaptation set was determined as 10% of the entire size of the neutral samples in the target set. While the size of the adaptation set was determined exploratory for the following experiments, similar research used 9% of the data from a target set (Abdelwahab and Busso, 2015). However, according to their research increasing the size (up to 35%) improved the classification results. On the other hand, increasing the size requires more human labour to annotate the unlabelled target data and thus extra cost.

Similarly, SVM classifiers with the linear kernel, SMO were trained from the adaptation set in addition to the training domain using 10-fold cross validation (It must be minded that during the cross validation, the selected neutral samples had been completely *excluded* from the testing sets). Thus, where D^p is the target domain, L=labelled data, U=unlabelled data and N is a set of labelled neutral reference:

Algorithm 1 Algorithm for adaptive binary SVM classifiers based on concatenating neutral references taken out from a target domain to the training data.

```
Let $D^t$ be the training domain
Let $D^p = D^p_{L_N} \cup D^p_U$, where $L_N \in \{N_1, N_2, \ldots, N_n\}$
f(x) classifier is trained on $D^p_{L_N} \cup D^t$, where f(x) is the SVM classifier
for $j \in \{1, \ldots, n\}$ do
    f(x) classifies $D^p_{U_i}$ to class 0 or class 1
end for
```

Results on Arousal Classification

Table 3.16 presents the classification results on arousal based on the precision and recall rates, using SVMs. The hand left side of the table contains results of the conventional cross-corpus approach, and the right hand side shows the results obtained from the proposed adaptive approach. Moreover, it must be noted that regarding the binary classification of arousal (low vs. high), neutral references were added to the class low (0) of arousal.

The mean values of the precision and recall rates of the conventional and adaptive proposed approaches were separately compared through paired t-tests. A repetitive pattern was observed in the results of arousal classification: The recall of class low (0), the class with the added neutral references, was increased (t= 3.3, p =.08). In addition, the precision of class 1, the class with no additional neutral data, was improved significantly (t=7.7, $p < .01$). Improvement in the obtained recall of class 0 implies the increase in the number of correctly classified instances under this class. In other words, the classifier has been able to detect a higher number of samples in this class. It should be noted that the class 0 of arousal constitutes of neutral and sadness classes for EMA, neutral, sadness and disgust for SAVEE and, neutral, sadness, disgust and boredom for the EMO-DB database.

In addition, Table 3.16 presents the G-Means of the two classes 0 and 1 in respect to the precision and recall metrics. The overall G-Means obtained for the precision was improved from 0.55 to 0.61 by adopting the adaptive proposed approach (t=2.34, p

			Conventional Ap.		Adaptive Ap.	
Training	Target	Class	Precision	Recall	Precision	Recall
IEMOCAP	EMA	0	0.66	0.46	0.57	0.83
		1	0.59	0.77	0.70	0.45
		G-Mean	0.62	0.60	0.63	0.61
	SAVEE	0	0.50	0.48	0.56	0.73
		1	0.51	0.54	0.64	0.45
		G-Mean	0.50	0.51	0.60	0.57
	EMO-DB	0	0.54	0.14	0.56	0.84
		1	0.50	0.88	0.67	0.33
		G-Mean	0.52	0.35	0.61	0.53
	Mean	0	0.57	0.36	0.56	0.80
		1	0.54	0.73	0.67	0.39
		G-Mean	0.55	0.51	0.61	0.56

Table 3.16: Precision and recall obtained from the Conventional and Adaptive proposed approaches in binary arousal classification 0 (low) and 1 (high). Grey horizontal rows present the G-Mean values of the recall and precision rates.

EMA		Conventional No. (Percentage %)	Adaptive No. (Percentage %)
High	Anger	**29** (17%)	95 (56%)
	Happiness	**52** (31%)	93 (53%)
Low	Neutral	**96** (56%)	36 (21%)
	Sadness	**79** (46%)	18 (11%)

Table 3.17: Number and percentage of misclassified instances in EMA database as the target domain by applying Conventional approach (left) and Adaptive proposed approach (right).

= .07). The overall G-Mean obtained for the recall metric was also improved from 0.51 to 0.56, however this increase was not significant either.

To explore in more detail how adding neutral references caused changes in the results of classification, the number and type of misclassified instances in each of the classes (0 and 1) were identified and counted. In particular, the number (as well as the percentage) of the misclassified instances in the subclasses of class 0 (such as neutral and sadness) and class 1 (such as anger and happiness) for arousal was examined.

Table 3.17, 3.18 and 3.19 respectively show the extent of misclassification for each emotion label for each pair IEMOCAP-EMA, IEMOCAP-SAVEE and IEMOCAP-EMO. From these tables, it can be observed that adding neutral references has calibrated the classifier to identify a higher number of instances from the neutral class. In addition to the neutral class, a higher number of samples from the sadness

SAVEE		Conventional	Adaptive
		No. (Percentage %)	No. (Percentage %)
High	Anger	29 (48%)	28 (47%)
	Happiness	26 (43%)	25 (42%)
	Frustration	25 (42%)	42 (70%)
	Surprise	32 (53%)	36 (60%
Low	Neutral	58 (48%)	22 (18%)
	Sadness	31 (52%)	18 (30%)
	Disgust	31 (52%)	22 (37%)

Table 3.18: Number and percentage of misclassified instances in SAVEE database as the target domain by applying Conventional approach (left) and Adaptive proposed approach (right).

EMO-DB		Conventional	Adaptive
		No. (Percentage %)	No. (Percentage %)
High	Anger	18 (14%)	82 (65%)
	Happiness	13 (18%)	47 (66%)
	Fear	10 (14%)	49 (71%)
Low	Neutral	69 (87%)	12 (15%)
	Sadness	47 (76%)	5 (8%)
	Boredom	71 (88%)	19 (23%)
	Disgust	44 (96%)	8 (17%)

Table 3.19: Number and percentage of misclassified instances in EMO-DB database as the target domain by applying Conventional approach (left) and Adaptive proposed approach (right).

and disgust classes have been correctly detected as well (in other words, a lower error rate can be observed in these classes).

As mentioned before, the precision of class 1 has significantly improved. One implication is that the number of falsely classified instances (False Positive (FP)) in this class (class 1) has been reduced. This implies reduction in misclassification of samples of neutral and sadness. According to the precision expression (3.2), where Precision is:

$$Precision = \frac{TP}{TP + FP}$$

For instance, the precision of class 1 of the EMA database would be:

$$Precision = \frac{No.Correct_{(anger,happiness)}}{No.Correct_{(anger,happiness)} + No.Error_{(neutral,sadness)}}$$

Precision by the conventional approach according to the results given in Table 3.17:

			Conventional Ap.		Adaptive Ap.	
Training	Testing	Class	Precision	Recall	Precision	Recall
IEMOCAP	EMA	0	0.54	0.71	0.54	0.32
		1	0.58	0.40	0.51	0.71
		G-Mean	0.56	0.53	0.52	0.48
	SAVEE	0	0.53	0.55	0.56	0.44
		1	0.52	0.51	0.54	0.66
		G-Mean	0.52	0.53	0.55	0.54
	EMO-DB	0	0.71	0.73	0.73	0.39
		1	0.21	0.19	0.26	0.59
		G-Mean	0.39	0.37	0.44	0.48
	Mean	0	0.59	0.65	0.61	0.38
		1	0.41	0.37	0.44	0.65
		G-Mean	0.49	0.49	0.52	0.50

Table 3.20: Precision and recall obtained from the Conventional and Adaptive proposed approaches in binary valence classification 0 (negative) and 1 (positive). Grey horizontal rows present the G-Mean values of the recall and precision rates.

$$Precision = \frac{(170 - \mathbf{29}) + (170 - \mathbf{52})}{(170 - \mathbf{29}) + (170 - \mathbf{52}) + (\mathbf{96} + \mathbf{79})} = 0.59$$

(Note that the number of instances in each class of the EMA corpus is 170)
Precision by the proposed approach:

$$Precision = \frac{(170 - 107) + (170 - 104)}{(170 - 107) + (170 - 104) + (36 + 18)} = 0.7$$

Consequently, concatenating a limited number of neutral references from the target domain to the training set suggests a significant increase on the recognition of neutral as well as sadness instances from the target domain.

Results on Valence Classification

Likewise, the conventional and adaptive approaches were applied for recognition on valence. Table 3.20 presents classification results based on the two measures precision and recall as well as their G-Means.

The means of the precision and recall rates of the conventional and adaptive proposed approaches were separately compared through paired t-tests. As it can be seen from this table, valence recognition through the proposed approach did not follow a behaviour similar to the arousal recognition. The recall of the class with the added neutral references (class 1) was significantly increased (t=3.92, p =0.05). However,

EMA		Conventional No. (Percentage %)	Adaptive No. (Percentage %)
Pos	Happiness	**117** (69%)	34 (20%)
	Neutral	**84** (49%)	22 (13%)
Neg	Anger	**37** (22%)	112 (66%)
	Sadness	**64** (38%)	122 (72%)

Table 3.21: Number and percentage of misclassified instances of valence classification in the EMA database as the target domain by applying Conventional approach (left) and Adaptive proposed approach (right).

on the contrary to the results of arousal classification, the precision of the other class (class 0) did not consistently improve across the different evaluations.

For instance, the precision of class 0 of the EMA database would be:

$$Precision = \frac{No.Correct_{(anger,sadness)}}{No.Correct_{(anger,sadness)} + No.Error_{(happiness,neutral)}}$$

Precision by the conventional approach according to the results given in Table 3.21:

$$Precision = \frac{(170 - \mathbf{37}) + (170 - \mathbf{64})}{(170 - \mathbf{37}) + (170 - \mathbf{64}) + (\mathbf{84} + \mathbf{117})} = 0.54$$

(Note that the number of instances in each class of the EMA corpus is 170)
Precision by the proposed approach:

$$Precision = \frac{(170 - 112) + (170 - 122)}{(170 - 112) + (170 - 122) + (22 + 34)} = 0.44$$

This could suggest that, although adding neutral references helped recognition of the rest of the neutral instances from the testing database, it had a less effective impact on the recognition of the happiness instances. This could be one of the reasons that the precision has been deteriorated in class 0 of valence.

Overall, adding neutral references from the target domain to the training set increased the correct recognition of the neutral instances for both arousal and valence. With respect to arousal, the adaptive models significantly improved the precision rate of the class high (class 1) by reduction of the FP rate (e.g. a lower number of misclassified instances of neutral to high arousal). Therefore, the adaptive approach

Corpus	Low (0)	High (1)
EMO-DB	Sadness, Fear Boredom	Anger, Happiness, Disgust
EMA	Sadness	Anger, Happiness
SAVEE	Sadness, Fear	Anger, Happiness, Surprise, Disgust
IEMOCAP	Sadness	Anger, Happiness, Frustration, Excited

Table 3.22: Mapping of discrete emotions of the EMO-DB, EMA, SAVEE and IEMOCAP corpora to ternary dominance. A separate class of Neutral (N) was considered for all the corpora.

appeared to be effective on the precision rate of the class high of arousal (class 1). On the other hand, the obtained higher accuracy could be linked to the fact that the adaptive training set is larger than the respective conventional set. However, the likeliness of this effect is weak due to this reason that in this experiment only 10% of the neutral samples were removed from the testing set and added to the training set.

With regard to valence, although the precision rate of the class low (class 0) was increased, this change was not significant. This result could suggest that the FP rate of the class negative (class 0) of valence was not reduced (e.g. the number of misclassified instances of neutral or happiness to negative valence was not decreased). This outcome could imply the plausibility of clustering neutral to the positive class of valence. However, more investigation is required to derive a solid conclusion.

3.4.2 Ternary Approach on Dominance Classification

The advantages of clustering emotion labels into binary valence, binary arousal and binary dominance on emotion recognition were discussed in Section 3.4. However, unlike valence and arousal dimensions, there is a lack of research on clustering neutral state into low- or high-class of dominance dimension. To tackle the neutral state, a separate class problem was introduced for dominance classification. In other words, the recognition task was addressed as a three-class problem, where the emotion labels were categorised into the low and high classes, and a separate class was considered for the neutral samples (Table 3.22).

As for all the other classification evaluations in this study, SVM classifiers with a linear function were used. However, this time SVMs were developed (trained and tested) for recognition of a three-class problem: Low, high and neutral. Likewise, the four emotional corpora EMO-DB, EMA, SAVEE and IEMOCAP were used to

Corpus	Dominance		
	Low (0)	High (1)	Neutral (N)
EMO-DB	212	198	79
EMA	170	340	170
SAVEE	120	240	120
IEMOCAP	470	600	400

Table 3.23: Number of instances in each ternary class of dominance: Low, High and Neutral.

Training	Target	UAR	G-Mean
EMA	IEMOCAP	0.47	0.46
	SAVEE	0.37	0.32
	EMO	0.32	0.06
	Average	0.39	**0.28**
IEMOCAP	EMA	0.41	0.29
	SAVEE	0.36	0.29
	EMO	0.27	0.15
	Average	0.35	0.24
SAVEE	IEMOCAP	0.41	0.14
	EMA	0.49	0.17
	EMO	0.37	0.00
	Average	0.42	**0.10**
EMO-DB	IEMOCAP	0.34	0.00
	EMA	0.42	0.00
	SAVEE	0.40	0.31
	Average	0.39	**0.10**

Table 3.24: Cross-corpus evaluation results on dominance based on unweighted average of class-wise recall (UAR) and geometric mean (G-Mean). The highest and lowest G-Means are shown in bold.

train four individual prediction models. Results of the classifications based on the class-wise recall and the total accuracy can be found in Table 3.24.

3.4.3 Classifiers from Multiple Corpora

Given the ternary classification of the dominance dimension and the inconsistent results obtained for binary arousal and valence recognition from the adaptive approach, this chapter concludes with the description of a set of approaches for measuring emotions based on the PAD dimensions. These approaches will be adopted in Chapter 5 (Main Study, Section 5.3.1) for measuring the emotional content of the thinking aloud verbalisations. It should be noted that the thinking aloud verbalisations constitute data in the testing/target set.

In this regard, an ensemble of classifiers will be used for analysing the PAD dimensions. An ensemble of classifiers (the idea of the ensemble of SVM classifiers has been proposed by Vapnik (2013)) is a set of classifiers whose individual decisions are combined in some way to classify the test instances. The main advantage of ensemble classifiers is that they usually result in a much higher accuracy than the individual classifiers that make them up (Kim et al., 2003; Schuller et al., 2011c). In this research, two ensemble algorithms were investigated to aggregate the results of classification from different SVMs ensemble.

Majority Voting Method: Majority voting (or plurality) is the simplest method for combining predictions from multiple classifiers. According to this method, each classification model contributes a single vote and the final prediction is assigned to the class with the most votes (see Algorithm 2) .

Algorithm 2 The majority voting algorithm to choose the most voted class for a given instance from multiple classifiers. The algorithm was adapted from Kim (2003)

```
1: Let $f_k$ denote a decision function of the $k$th classifier, where k $\in \{1, 2, \ldots, K\}$
2: Let $c_j \in \{0, 1\}$ denote a label of the $j$th class, where j $\in \{1, 2\}$
3: Then, let $N_j$=$\#\{k|f_k(x)= c_j\}$ is the number of classifiers whose decisions are
   known to the $j$th class
4: Then, the final decision $f_D(x)$ for a given test vector $x$ from the majority voting
   is determined by: $f_D(x)$=arg$_j$ max$N_j$
```

Majority voting assumes that all the classification models are equally good to estimate a label. However, for example, if there is only one true optimised classifier, and the majority are not, and if unoptimised classifiers give the same incorrect label to a specific instance, then the majority voting method would favour them since they are in a majority. One could address this problem by introducing a weight capturing how good each classifier is.

Weighted Voting Method: In weighted voting (Kuncheva and Rodríguez, 2014), individual classifiers have various degrees of influence on the prediction, in which their influence is proportional to their assigned weight. Accordingly each classifier is associated with a specific weight that can be determined by a number of factors such as the classification accuracies, errors and the degree of sensitivity or specificity of their prediction results. In general, the final prediction is decided by summing all the weighted votes and by choosing the class that has the highest aggregate.

Adopting the weighted voting technique, I propose a weight coefficient relevant to the size of the corpus (given the use of the three English corpora with different distributions). To determine the weights, it was assumed that the distribution of a class in a corpus implies the power of generalisability of the corpus for prediction on future data of that specific class.

Thus a class of c_j label of a corpus with a larger distribution receives a higher weight than the same class of c_j in another corpus with a lower distribution. Therefore the weight coefficient for the class c_{ji} with distribution N_{ji} in training set T_i could be obtained from the Equation 3.3. In this equation, if the weights are normalised, their sum will be equal to 1.

$$w_{ji} = \frac{Nji}{\sum_{j=1} N_j} \tag{3.3}$$

Algorithm 3 The majority weighted voting for binary decision.

```
Let $f_k$ denote a decision function of the $k$th classifier, where k $\in \{1, 2, \ldots, K\}$
Let $c_j \in \{0, 1\}$ denote a label of the $j$th class, where j $\in \{1, 2\}$
Let $w_k \in$ [0,1] and $\sum_{k=1} w_k$= 1 denote a weight of the $f_k$
Then, the final decision $f_D(x)$ for a given test vector $x$ from the weighted voting
  is determined by:
if $\sum_{j=1} w_k \times c_j >= 0.5$ then
    The final decision $f_D(x)$= $c_2$
else
    The final decision $f_D(x)$= $c_1$
end if
```

Arousal Recognition

According to the process of the proposed adaptive approach (Section 3.4.1), the neutral references from each individual are concatenated to the related class of the IEMOCAP corpus. As it was shown in the results of arousal classification by adopting this approach, the adaptive approach led to a significant improvement in the precision of the class high (class 1) of arousal (reaching 70% precision). This result implies that the target data being predicted as class 1 label, have been correctly classified with a high probability (due to improvement achieved in the precision rate). Hence the target data labelled with 1 is marked as correctly classified and subsequently removed from the rest of the analysis.

On the other hand, the target data labelled with class 0 is considered for the second round of classifications. In this regard, the classification models are created using SVM algorithms and based on the three English corpora and according to the binary approach for the low (0) and high (1) classes of arousal. To aggregate the results of classification from different SVM ensembles, the majority weighted voting for binary decision Algorithm (3) will be employed in Section 5.3.1.

Valence Recognition

As it was shown the adaptive classification approach did not result in any significant improvement for recognition on valence. Therefore the majority weighted voting for binary decision Algorithm (3) will be applied to aggregate the results of different classifiers derived from the multiple corpora in Section 5.3.1.

Dominance Recognition

Likewise, I propose a method similar to the weighted approach to analyse the dominance dimension. In this regard, given the three-class problem of dominance, the following three labels are considered: Low (0), neutral (0.5) and high (1). Next, using the Equation 3.3, weights related to each classifier are obtained. The procedure of determining a final decision label for the dominance dimension is shown in Algorithm 4.

Algorithm 4 The majority weighted voting for ternary (or military) decision.

```
1: Let $f_k$ denote a decision function of the $k$th classifier, where k $\in \{1, 2, \ldots, K\}$
2: Let $c_j \in \{0, N, 1\}$ denote a label of the $j$th class, where j $\in \{1, 2, 3\}$
3: Let $w_k \in$ [0,1] and $\sum_{j=1} w_k$= 1 denote a weight of the $f_k$
4: Then, let $W_j$=$\{\sum_{j=1} w_k \mid f_k(x) = c_j\}$ is the number of classifiers whose decisions
   are known to the $j$th class
5: Then, the final decision $f_D(x)$ for a given test vector $x$ from the greatest vote is
   determined by: $f_D(x)$=arg$_j$ max$W_j$
```

3.5 Summary

This chapter presented the overall requirements and components of the emotion recognition systems based on vocal expressions. Specifically two approaches were described: The first approach was carried out by developing the classifiers trained on one corpus, the IEMOCAP database (Section 3.3). The second approach was followed by developing ensemble of classifiers from multiple corpora and employing decision-fusion methods to combine their class predictions (Section 3.4).

The classifiers developed through the first approach will be used for prediction on the emotional content of the thinking aloud utterances collected in the Pilot Study, the next Chapter.

Chapter 4

Study 1: Vocal Emotional Analysis of Think Aloud Verbalisations

An exploratory study was conducted for measuring emotional aspect of UX by analysing expressive voice. To elicit emotions, an online shopping application offering travel recommendations was created. The think aloud technique was carried out as a means for collecting concurrent emotional experiences through vocal expressions. Retrospective and subjective assessments of emotions were carried out by employing self-report questionnaires. Results of the study confirmed the previous finding that retrospective assessments did not correspond to the actual experience. The results also suggested that retrospective assessments of emotions were significantly correlated with the most frequently elicited emotion (i.e. modal emotion) during the interaction.

Finally, in addition to emotional aspect of UX, other qualities such as aesthetic and usability were assessed by self-report questionnaires at the end of the study. The results confirmed the findings of previous studies, where the perceived design quality of an application, which was also measured after the interaction with the application, was significantly correlated with retrospective assessment of emotions.

4.1 Aim of the Study

The main objective of this study was to examine the effectiveness of the think aloud in combination with ML techniques for concurrent and automatic assessment of vocal emotions in the context of UX.

As it was explained in Section 2.3.2 the retrospective evaluations are often taken as the indicator of the entire experience. The underlying assumption is that retrospective evaluations can reflect the average or sum of all the moments encountered during the interaction, although the research has revealed that this assumption does not always hold.

To enhance the repertoire of approaches used for emotion analysis, I argue for the think aloud technique and vocal emotional analysis of the verbalised content, which can be used to derive people's emotions without requiring any sophisticated procedure or equipment. Another advantage of this approach is collecting data elicited from motor expressions, where the data can be used for triangulating with subjective self-reports and to obtain a more comprehensive view of emotional UX.

To examine the effectiveness and validity of the proposed methodology, an exploratory study was designed. This study was based on the two temporal stances of measurement (concurrent vs. retropsepctive) and therefore the associated questions such as the peak-end effect. Consequently, two hypotheses were formulated and investigated on the basis of the following two concepts: Kahneman's *Total Experience* and Scherer's *Modal Emotions*.

Total Experience: According to Kahneman and colleagues Kahneman et al. (2003), an experience constitutes moments of pleasure and pain, and total experience of an episode is the aggregation of these moments. The retrospective evaluation of an experience is prone to the peak-end effect, hence, a correct assessment of total experience can be permitted only by a moment-based approach.
Given this concept, it was assumed that each utterance given during thinking aloud serves as a moment of experience, which can be attributed by the three dimensions of emotion (the PAD framework Mehrabian and Russell, 1974). In this regard, the total experience of an interactive session can be derived from integrating the emotional values of all utterances collected real-time from a particular participant. Therefore, aligned with Kahneman's concept of memory and moment-based assessment of experience, I argue that the retrospective assessment of experience (measured by the PAD dimensions) does not reflect a correct estimate of total experience and formulated hypothesis (H) as follows:

H1: Retrospective evaluation of the PAD dimensions (**H1a**: Pleasure; **H1b**: Arousal; **H1c**: Dominance) will not be significantly correlated to the total experience of PAD during the actual interaction as derived from vocal emotional analysis.

It should be noted that, since the three dimensions of PAD are known to be independent (e.g. Mehrabian, 1995; Yik et al., 1999), their corresponding values were considered to be aggregated separately.

Modal Emotions: According to Scherer (2005), there are a small number of emotions that are relatively frequently experienced such as anger and joy. Scherer suggested calling these emotions Modal Emotions. In a think aloud session, the modal or most frequently elicited emotion can be obtained by vocal emotional analysis and by taking the discrete model of emotions.
The following hypothesis was formulated to investigate to what extent the modal emotion identified by vocal analysis can reflect the retrospective assessment of emotions:

H2: The more positive the modal emotion is, the more positive the retrospective appraisal of the experience is; the same relation is hypothesised for a negative modal emotion.

4.2 Experimental Design

4.2.1 Participants

Forty six participants (female=25, male=21, with the average age of between 18-25) voluntarily took part in this study. Thirty eight of the participants were native English speakers and the rest were very fluent in English. The participants were invited through personal contacts. The sample consisted of university students from the undergraduate (n=26) to postgraduate (n=17) level and 3 participants were university academics or administrative staff. Most of the students participating in this study majored in computer science, followed by geography, law, physics, management, and chemistry. However, data of 5 participants were excluded from data analysis, because they were either incomplete or invalid (e.g., participants failed to follow the instructions).

4.2.2 Instruments

E-commerce Application

An e-commerce of selling travel services was chosen as the context for this study. Online purchase of travel services is rapidly growing due to the ability of the Internet to provide travel customers with a variety of options, often with reduced prices, and having a convenient way to shop (Anckar and Walden, 2001). Travel products are experience goods , meaning that their quality cannot be determined prior to the purchase and the actual experience (Peterson et al., 1997). As a result, purchasing travel services is characterised as a problem-solving process due to the need for acquisition of knowledge of the holiday destination (Moutinho, 1987). On the other hand, depending on the purposes of the purchase, travel booking can be associated with hedonic qualities such as pleasure, stimulation and enjoyment (Holbrook and Hirschman, 1982). Furthermore, resulting from the anticipated interaction (Sward and Macarthur, 2007), certain subjective experience can be elicited before the actual interaction with the travel booking website. For instance, when one has heard or experienced about a specific travel agency or destination, the expectation could shape before-interaction experience.

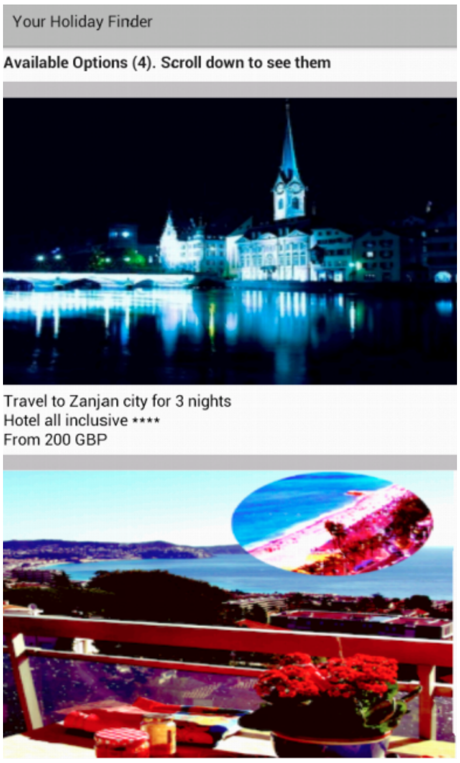
In the design of online travel systems, images play an influential role in consumers' decision-making processes and behaviours (Rezende-Parker et al., 2003). Images construct a mental impression of the attributes and benefits gained from a destination. Moreover, destination images have been served as emotional stimuli in addition to their role of cognitive engagement (Baloglu and Brinberg, 1997). Therefore, in the design process of this study's application, I put a special emphasis on providing multiple images to evoke emotional responses.

In total, 20 images of different locations were evaluated in terms of their perceived hedonic qualities such as identification, stimulation, and attractiveness using the Attrakdiff questionnaire. Seven people (3 female and 4 male), who were not the participants of the study, assessed the images.

Other design aspects that have been considered include the presentation of appropriate information, registration and check-out, displaying cheap services and having confirmation page (Safavi, 2009). Given the above design criteria, I developed an Android-based application, called HolidayFinder as an online shopping platform that offered discount travel recommendations. In this application, the selection of each product (the travel recommendations) was processed by three subsequent interfaces. The first interface contained details of the product, including multiple images and textual contents. The second interface included the checkout form to be filled in by the user. Finally the last interface contained the "Thank You" message and a button to continue shopping. Screenshots of different interfaces in HolidayFinder are illustrated in Figures 4.1 and 4.2 .

Audio-recording Application

I also created a Java-based application to record the audio data captured by the headset microphone at 44.1 kHz, 16bits as a wav file for each participant. This is the standard configuration that was used by the openSMILE software to process audio data.

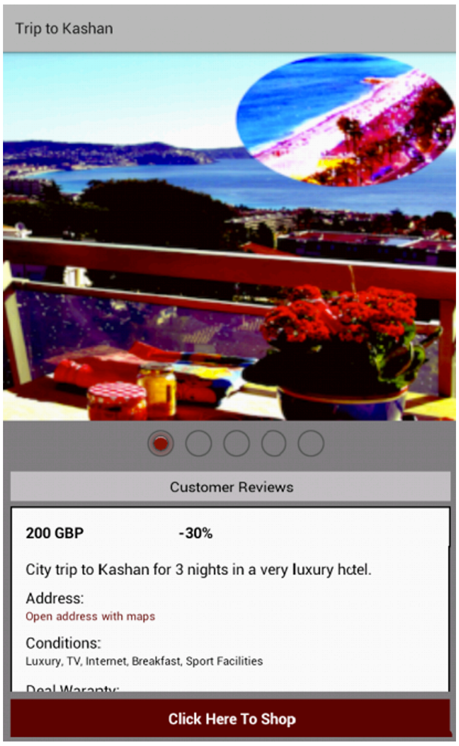


Your Holiday Finder

Available Options (4). Scroll down to see them

Travel to Zanjan city for 3 nights
Hotel all inclusive ****
From 200 GBP

Figure 4.1: Screenshot of the first interface of HolidayFinder.



Trip to Kashan

200 GBP -30%

City trip to Kashan for 3 nights in a very luxury hotel.

Address:
Open address with maps

Conditions:
Luxury, TV, Internet, Breakfast, Sport Facilities

Deal Waranty:

Figure 4.2: Screenshot of the second interface of HolidayFinder.

Hardware

The hardware that was used in this study included a tablet, 7-inch screen size, with the Android operation system and a laptop running Windows 7 equipped with an external microphone headset to capture the audio input and to filter out ambient noise.

Porat and Tractinsky Scale

This scale was used to measure the self-reported perception of the interaction, which is based on the Porat and Tractinsky research model of UX as described in 2.1.1. To reiterate, this model proposed that an environment (such as a product/service, e.g. websites), due to its characteristics (such as the usability and aesthetics qualities), induces emotional reactions in individuals, and subsequently the emotional reactions influence the attitude toward the environment.

The internal consistency of all the sub-scales were calculated using Cronbach's alpha coefficient of reliability (represented in Table 4.1).

	1	2	3	4	5	6	7
1. Pleasure	0.87						
2. Arousal	0.38	0.75					
3. Dominance	0.64	0.64	0.63				
4. Classical aesthetics	0.48	ns	0.64	0.85			
5. Expressive aesthetics	0.54	0.37	0.64	0.65	0.86		
6. Usability	0.31	0.38	0.64	0.70	0.56	0.82	
7. Attitude	0.52	0.36	0.64	0.83	0.50	0.72	0.96
8. Modal emotion	ns	ns	ns	ns	0.28	ns	0.27

Table 4.1: Correlations between the constructs of Porat and Tractinsky scale. Grey diagonal cells show Cronbach's alpha coefficient of reliability.

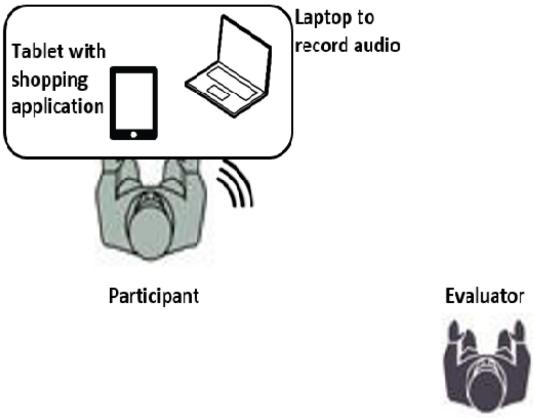


Figure 4.3: Experimental setup.

4.2.3 Procedure

The study ran as a series of lab-based sessions on an individual basis. In the beginning of each session the objectives of the study was explained. Then the procedures of the study, including the introduction to the think aloud protocol was described. Next, the session continued by a short think aloud trial phase where the participants were instructed to verbalise thoughts and feelings while interacting with the Amazon.com website. In this regard, participants were reminded that during thinking aloud, they would be the primary speakers or contributors and that the experimenter would primarily play a listener role (Boren and Ramey, 2000). The inclusion of the initial practice accords to the Ericsson and Simon's framework, where participants should practise thinking aloud before the test begins (Ericsson and Simon, 1993).

After the introduction phase, the actual experiment began. In this phase, the participants were asked to complete a shopping process according to the following scenario and talk out loud their thoughts and emotions during their interaction, which were recorded by the java-based audio recorder.

> *Your friend's wedding is coming in three days from today and you need to get him/her something as a gift. You are aware of his/her interest in taking pleasurable trips. There is an application in your tablet called HolidayFinder, which offers different holiday options. You find today a good time to buy a gift for your friend.*

During the interaction phase, the experimenter minimised her interference by acknowledging the verbalisations through the use of tokens such as "yeah" and "uh-huh" and prompting participants to keep on thinking aloud if they remained silent for 15 seconds.

Upon the completion of the given task, the participants were asked to fill out the final questionnaire, which consisted of the Porat and Tractinsky scale. Finally at the end of the session, participants were given time to provide optional feedback about the entire experiment and then they were thanked for their participation. Each participant was allocated a one-hour session. However, most of the sessions were completed before the planned time.

4.2.4 Segmentation

Transcriptions of what has been said is a necessary task for further analysis. Automatic speech-to-text-recognition (STR) can facilitate this process. However, the problem with the STR technology is the low reliability and accuracy of the existing tools. For instance, some common STR tools have achieved less than 40% of accuracy (Lasecki et al., 2017) and other available tools with higher accuracy require intensive training (Kawas et al., 2016; Shadiev et al., 2014). Therefore, many studies in voice emotion recognition have relied on the manual annotation of spoken data to eliminate additional errors that can be introduced by STR tools.

Consequently, the manual transcription and segmentation of the audio recording into meaningful utterances was chosen for this study. Transcription is a relatively straightforward, albeit time consuming task. On the other hand, segmentation is not straightforward because it is difficult to ensure that the same emotional state is maintained throughout an utterance.

A two-step procedure was used for the manual segmentation. First the audio recording were chopped into chunks using the Audacity software and based on the silence

Utterances	Labels	P	A	D
It seems relatively straightforward to book.	Happiness	4	4	3
I can not open the address.	Sadness	2	4	2
Ah..Probably it looks a bit warmer.	Happiness	3	2	2
But you can't really see clearly what it is.	Anger	2	4	2

Table 4.2: Example of transcription of utterances and corresponding recognised emotions by acoustic analysis. Note that the dimensions P (pleasure), A (arousal) and D (dominance) were recognised on the scale 1 (low)- 5 (high).

pauses of longer than 1 second. Next, two raters were involved to evaluate each chunk as a meaningful unit with the same emotional state (utterance). If the chunk was evaluated as an utterance, the value of 1 was assigned, otherwise 0. In addition, to assist with a clearer idea of what it was meant by a meaningful utterance, examples of earlier segmented transcriptions from the IEMOCAP corpus were given to the raters.

In a number of studies such as the INTERSPEECH 2009 Emotion Challenge, the utterance-level for emotion analysis (2.66 words per utterance on average) have been used. The IEMOCAP corpus, which was used as the training set in this study, was also segmented based on the utterance-level (with the average of 11.4 words per utterance).

In this process, regular discussions were held to negotiate the discrepancies in segmentation. As a result, some incomplete sentences such as "it seems..." were disregarded. Moreover some multiple-sentence utterances were split into shorter chunks, where the two raters agreed upon.

One of the raters was myself (an HCI researcher) and the other was an electrical engineer who is familiar with HCI work in general but was not involved in designing or running the empirical studies.

The final level of agreements in the segmentation task was measured by Cohen's Kappa (Fleiss and Cohen, 1973). The value of Kappa was 0.74, which is a reasonable level of agreement (Robson, 2002). In total, 1499 utterances (mean=39.4, SD=24.4 utterance per participant) were manually defined and segmented (10.31 words per utterance on average).

4.2.5 Automatic Recognition: Acoustic Analysis

Using the IEMOCAP corpus as the training dataset, and following the process of emotion recognition using the SVM algorithms (Section 3.3), the emotional state of each utterance was detected based on the discrete and dimensional models of emotions. It should be noted that, prior to recognition, the acoustic features of each segment was extracted using the openSMILE and INTERSPEECH 2009 feature set.

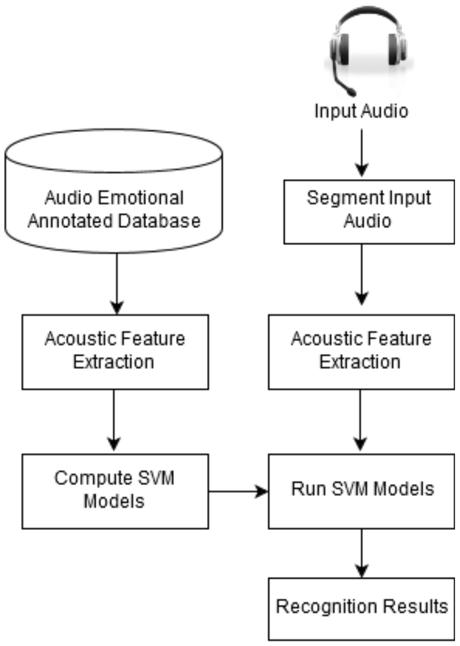

Input Audio
Audio Emotional Annotated Database
Segment Input Audio
Acoustic Feature Extraction
Acoustic Feature Extraction
Compute SVM Models
Run SVM Models
Recognition Results

Figure 4.4: A flowchart of the functionality implemented for emotion recognition from acoustic features.

The overall procedure of emotion recognition using the machine learning algorithms is illustrated in Figure 4.4.

As a result of recognition, for each utterance one emotion label (labels such as happiness, frustration, anger, sadness, and neutral) and three integer numbers (1: low, 5: high) corresponding to the level of pleasure, arousal and dominance were detected (Table 4.2).

4.2.6 Manual Recognition: Human Validation

To check the validity of the recognised emotion values, the same two raters for the segmentations task were involved to assess the emotional attribute of each utterance based on the three dimensions of emotion - pleasure, arousal and dominance - on a 5-point interval scale, ranging from 1(low) to 5 (high). The Kappa value was computed to test the reliability of the evaluations between the two raters: Kappa pleasure= 0.51, Kappa arousal= 0.77 and Kappa dominance= 0.53, while the Kappa scores between 0.4- 0.6 are considered to be fair (Robson, 2002). The rather low values of Kappa for pleasure and dominance can be attributed to the fuzziness in identifying emotions (D'mello and Graesser, 2007).

4.3 Data analysis

4.3.1 Normalising Total Experience

As explained previously, to obtain a value of total experience, the corresponding values of the PAD dimensions for all utterances were summed for each of the participants. However, due to individual differences in think aloud verbalisations, the number of utterances varied significantly from one participant to another. To control this, the value of total experience was averaged based on the total number of utterances for each participant. In other words, for all participants, the value of dimensions- pleasure, arousal and dominance - were normalised to a number in a range from 1 to 5.

Therefore, where N is the total number of utterances for the participant j:

Algorithm 5 Algorithm to obtain a numerical value for Total Experience

```
for $dimension \in \{P, A, D\}$ do
    for $i \in \{1, \ldots, N\}$ do
        $(\text{Total Experience})_{\text{dimension}} = (\text{Total Experience})_{\text{dimension}} + \text{Value}_{\text{dimension i}}$
    end for
    $(\text{Normalised Total Experience})_{\text{dimension}} = (\text{Total Experience})_{\text{dimension}} / \text{N}$
end for
```

4.3.2 Quantifying Modal Emotion

According to the circumplex model of affects (Russell, 1980), emotions can be described in a two-dimensional space defined by pleasure (P) and arousal (A). Thus, each emotion can be understood as the product of these two independent dimensions. To obtain a comparable measure between the retrospective assessment of emotions described in terms of the PAD dimensions and the modal emotions detected by vocal analysis, the following procedure was performed for each participant:

First, the different emotion labels, which were detected by vocal analysis and assigned to the utterances of each participant, were counted. For instance, for a $\text{participant}_{\text{j}}$, 20 times anger, 5 times sadness, 7 times happiness and 16 times neutral were detected. Among the labels, the most frequent occurring emotion was identified (e.g. anger was the most frequent emotion label for this particular participant).

Subsequently, across all participants, the modal emotion was derived: Anger was the modal emotion for most of the participant (N=23), neutral was in the second place (N=13), followed by happiness (N=5).

To evaluate the relationship between the modal emotion detected from the acoustic analysis and the corresponding retrospective measure of emotions, a numeric value was assigned to each type of modal emotion: Happiness was assigned to 3 (most positive detected emotion), neutral to 2 and anger to 1 (most negative detected emotion). For instance, if for $\mathrm{participant}_\mathrm{j}$ the modal emotion was anger, value 1 represented this emotion category. This designation was inferred from the Core Affect (Russell, 2003) of emotions, where emotion labels are placed in a circular order and neutral state at the centre point.

At the end, the score $E{=}P{*}A$ from the self-reported retrospective assessments of pleasure and arousal dimensions was obtained. The score E was considered as the corresponding measure of the modal emotion derived from the retrospective assessment. It was expected to find a pattern that the highest scores of P*A refer to the positive modal emotion, which was happiness, the lowest scores of P*A refer to Anger and the medium scores of P*A refer to Neutral as the modal emotion.

4.4 RESULTS

4.4.1 Recognition Performance

Manual human ratings were used as a benchmark against which the automatic recognition were validated. Specifically, the overall accuracy for each of the three dimensions Pleasure, Arousal and Dominance was calculated. Accuracy is the total number of correctly classified samples (utterances) divided by the total number of samples (utterances) to be classified in the testing set (thinking aloud data). Accuracy can also be referred to as the weighted average of class-wise Recall rates. As results, the obtained accuracy for each of the Pleasure, Arousal and Dominance was 0.30, 0.27 and 0.34 respectively. It should be noted that all the recognitions were based on 5 classes. As it can be seen, the results suggested a higher accuracy for Dominance as compared to Pleasure and Arousal.

4.4.2 Descriptive Statistics

PAD Dimensions

Table 4.3 presents the mean, standard deviation and max values of the PAD dimensions taken from the retrospective questionnaire (which is designated as PAD_R), moment-based analysis of PAD collected during the experience and assessed using the acoustic-emotion analysis (which is designated as PAD_Ac) and manually by the human raters (which is denoted as PAD_M).

	P_R	P_{Ac}	P_M
Mean	3.31	2.14	2.22
SD	0.75	0.33	0.23
Max	4.80	4.00	3.60
	A_R	A_{Ac}	A_M
Mean	3.14	4.22	2.1
SD	0.62	0.44	0.35
Max	4.25	3.95	2.7
	D_R	D_{Ac}	D_M
Mean	3.33	2.01	2.25
SD	0.55	0.14	0.26
Max	4.42	2.35	2.89

Table 4.3: Descriptive analysis of PAD dimensions taken from retrospective questionnaire and moment-based emotion analysis assessed automatically using acoustic features and by human evaluators. The subscript R denotes Retrospective, Ac denotes Acoustic analysis and M refers to Manual assessment by the human raters.

Three Friedman tests (Friedman, 1940) were performed; One for each of the three dimensions- with Pleasure, Arousal and Dominance being the dependent variables (DV) and the type of assessment method (Retrospective questionnaire, automatic Acoustic analysis and Manual assessment by human raters) as the independent variable (IV).

In the measures of Pleasure, there was a statistically significant difference between the assessment methods, $\chi^2(2) = 20.32$, $p < 0.002$. Post hoc analysis with Wilcoxon signed-rank tests was conducted to examine the differences. Results showed significant differences between Retrospective and Manual assessment (Z = -3.19, $p < 0.001$), and between the Retrospective and Acoustic assessment (Z = -3.19, $p < 0.001$). However there was no significant difference between the Acoustic and Manual assessment (Z = -1.16, p = 0.248).
With respect to Arousal, there was also a statistically significant difference between the assessment methods, $\chi^2(2) = 26.00$, $p < 0.001$. Post hoc analysis with Wilcoxon signed-rank tests revealed the following results: There were significant differences between Retrospective and Manual assessment (Z = -3.18, $p < 0.001$), between the Retrospective and Acoustic assessment (Z = -3.18, $p < 0.001$) and between the Acoustic and Manual assessment (Z = -3.12, p = 0.248). The mean scores of Arousal measured by the three approaches were in this order $A_{Ac} > A_R > A_M$.
For Dominance, likewise a statistically significant difference was seen $\chi^2(2) = 14.00$, $p < 0.001$. Results of the post hoc analysis showed significant differences between Retrospective and Manual assessment (Z = -2.97, $p < 0.01$), and between the Manual

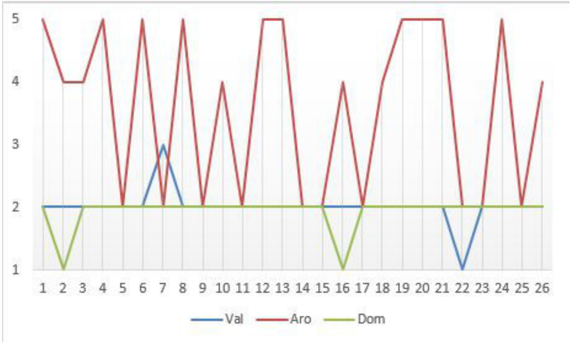


Figure 4.5: Example of emotional changes profile measured by acoustic analysis (participant 2).

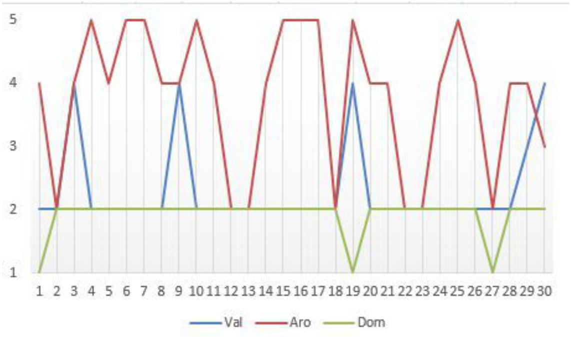


Figure 4.6: Example of emotional changes profile measured by acoustic analysis (participant 21).

and Acoustic assessment ($Z = -3.11$, $p < 0.01$). However, no significant difference was shown between the Retrospective and Acoustic assessment ($Z = -.382$, $p = 0.701$).

The significant differences between the momentary (Acoustic and Manual) and retrospective assessments in measuring the PAD dimensions could indicate the memory-experience gap. Especially, the higher values of $PAD_R > PAD_M$ in respect to all three dimensions could imply that retrospective ratings are the overestimations of the actual experience (Miron-Shatz et al., 2009).

In addition, Figures 4.5 and 4.6 represent two participants' emotional changes during their interaction measured by acoustic analysis. They show that emotional arousal changes more quickly over time than emotional valence and dominance.

Perceived Design Dimensions

Table 4.4 shows the mean and standard deviation of the perceived design dimensions in terms of aesthetics and usability qualities in a 7-point scale. Overall, the par-

	Classical aesthetics	Expressive aesthetics	Usability
Mean	4.29	3.12	5.20
SD	1.37	1.35	1.25

Table 4.4: Mean and standard deviation (SD) of classical aesthetics, expressive aesthetics, and usability measured by Porat-Tractinsky 7-point scale.

ticipants rated the HolidyFinder application as above average in terms of classical aesthetics and usability and below average in terms of expressive aesthetics.

4.4.3 Correlations

To evaluate H1 and H2, the following statistical analysis were performed. A Spearman's rank-order correlation was run to determine whether associations exist between the measures taken from the three assessment methods, including the use of questionnaire, acoustic-emotion analysis and human ratings in evaluating the PAD dimensions. Results of the analysis showed no significant correlation in the measures of the Pleasure and Arousal dimensions. However, positive and significant correlations were found for the Dominance dimension. The results showed that there were significant correlations between $\mathrm{D_{Ac}}$ and $\mathrm{D_R}$ ($\mathrm{r_s}(39) = .432$, $p = .02$) and between $\mathrm{D_M}$ and $\mathrm{D_R}$ ($\mathrm{r_s}(39) = .57$, $p = .04$). Therefore, we concluded that H1a and H1b were accepted, and H1c was not supported.

Furthermore results of a Spearman's rank-order correlation test between the modal emotion and the *P*A* score derived from the retrospective assessment showed a statistically significant positive correlation ($\mathrm{r_s}(39) = .339$, $p = .05$). Therefore, H2 was accepted.

Additionally, Table 4.1 shows the significant correlations between the perceived design attributes, self-reported emotions and attitude toward the application ($p = .01$). These results confirm previous findings, where the perceived qualities of an application are significantly associated with the overall perceived emotions (Porat and Tractinsky, 2012).
Moreover, weak correlations were found between modal emotion and expressive aesthetics ($\mathrm{r_s}(39) = .28$, $p = .086$) and between modal emotion and the attitudes toward the application ($\mathrm{r_s}(39) = .27$, $p = .95$).

4.5 Discussion

4.5.1 Implications of the Study Design

Product category

Travels booking was the choice of domain for the design of the e-commerce application. The literature review suggests that this service category can induce hedonic as well as utilitarian attribution in consumers' behaviours (Crowley et al., 1992). For example, services such as recommending holiday destinations, expensive restaurants and products like music-related appliances most likely would land in hedonic categories, whereas products such as calculators and cars would fall into the utilitarian category (Crowley et al., 1992). To avoid such influences, products that can be conceived equally hedonic and utilitarian would be suitable for future work, since the focus of the study was not to evaluate the effects of the e-commerce domain.

Think Aloud Protocol

The use of think aloud protocol was proposed as a method for collecting nonverbal data in addition to verbal data during UX test sessions. In the case of usability research, verbalisations as well as behavioural data are collected for analysing the cognitive demand of a system such as identifying deficiencies (Boren and Ramey, 2000). Moreover, some researchers such as Dumas and Redish (1999) suggested using questions to ask how participants feel to probe preferences and explanations during thinking aloud (Rubin and Chisnell, 2008). Nontheless it was observed that often usability engineers ask participants to think aloud about what they like or what they do not like (Boren and Ramey, 2000). These practices implicitly suggest the use of the think aloud protocol for gathering not only the cognitive information, but also experiential tendencies as well. This type of verbal protocol is regarded as relaxed think aloud (as it was explained in 2.2.1), which allows for probing participants for their feelings and opinions (Nielsen et al., 2002). Consequently, it deemed plausible to propose that beside the verbal data, behaviours and expressions manifested during the system interaction could be used for analysing emotional experiences.

In this study, the think aloud protocol was based on Boren and Ramey's speech communication theory (Boren and Ramey, 2000), which defines the role of a speaker, who interacts with and thinks aloud about an interface, and that of a listener, who is there to actively listen to the information. On the other hand, IEMOCAP is a corpus designed for recognition of emotions in dyadic communications. Therefore, employing the speech communication as a practical alternative of the original think

aloud protocol by Ericsson and Simon (1993) matched the prerequisite of using the IEMOCAP corpus for detecting emotional expressions.
However, despite the communication nature of the employed think aloud technique, the recommendations of Ericsson and Simon (1993) model were taken into account: First, before the actual test, detailed instructions were given and the initial practice was performed. Second, the interaction between experimenter and participant kept minimised. Finally, the reminder "keep talking" was used to prompt participants for talking after remaining silence for 15 seconds. This reminder is considered as non-directive and short, at the same time reflects the presence of the experimenter.

4.5.2 Memory-based vs. Moment-based Assessments

In comparison between retrospective and moment-based assessments of the PAD dimensions, one significant correlation was found; positive correlation between D_{Ac} (moment-based evaluation of dominance through acoustic emotional analysis) and D_{R} (retrospective evaluation of dominance by the use of questionnaire). According to Kahneman (Kahneman et al., 2003), experience is characterised by the moments of good or bad (pleasure or valence) with different intensity from mild to extreme (arousal). Based on this conception, dominance or the level that a person feels in control of situation should not be considered as an attribute of the moment of experience. In addition, the reason for finding the significant correlation between D_{Ac} and D_{R} can be explained by the assumption that dominance is likely more stable over time, not as fluctuating as pleasure and arousal, which change from moment to moment (cf. Figures 4.5 and 4.6). As a result, the retrospective evaluation of the dominance dimension is correlated with that of the moment-based assessment reflected about this dimension. In addition, previous research has also argued that dominance reflects a more cognitive reaction and less of an emotional state (Barrett and Russell, 1999). This is another argument that can explain why there is no significant difference between memory-based and moment-based evaluations of dominance.
Previous research has shown that retrospective evaluations of past experiences are influenced by two moments, namely: the moment with the highest intensity (i.e. the peak) and the final moment (i.e. end) (Kahneman, 2011; Kahneman et al., 1993). Results of this study suggested that the overall evaluation of an experience could be predicted by the most frequently elicited emotion (i.e. modal emotion) during the experience. It can be argued that the most frequently evoked emotion has become the most remembered emotional response for an event (cf. the notion of rehearsal in the theory of memory Craik and Lockhart (1972)), and therefore this modal emotion correlates strongly with the retrospective evaluation, which also

relies on memory. This observation is also linked to Kahneman's Kahneman (2011) arguments on 'experiencing self' versus 'remembering self', which in turn is closely related to psychological well-being. Nonetheless, the intricate relationships of these constructs entail more research efforts.

4.5.3 Recognition of Vocal Expressions

The rather low recognition rates reveal the complexity of the process as well as the shortcomings of the current analysis. The IEMOCAP database was used for training the recognition models. Despite the large number of instances included in this database, the context of the recordings is not UX related. In other words, the chosen scenario or context has an impact on the range and intensity of the elicited emotions.

On the other hand, the obtained results implied that a learner model can still be trained to predict certain emotions, and in particular for the dominance dimension in this study. Previous research has also shown that accuracy is highly dependent on the specific sub-category of emotions considered Schuller et al. (2010b). Thus, it is important to adopt strategies (such as employing an adaptive approach), which can cope better with the difficulties related to real-life data.

4.5.4 Threats to Validity

Threats to internal validity concern confounding factors that can influence the findings (Campbell, 1957). Results of the study showed no significant correlations between human ratings and automatic analysis on the vocal recognition of emotions. This can stem from the underlying issues in both approaches. Discriminating between 5 levels of emotional states for each dimension, judged by human raters, can make the ratings susceptible (D'mello and Graesser, 2007). In particular the lower Kappa scores (Section 4.2.6) in valence and dominance ratings can corroborate the high subjectivity and ambiguity of these ratings. Additionally, the issue of interpreting emotions from relying only on vocal cues could have been reinforced due to lack of information usually gained in face-to-face encounters.

Similar discrimination problem can be traced from the results of automatic analysis, such as the rather poor classification accuracy. In addition, distinguishing neutral states from other emotional states can cause uncertainty. Therefore, no verified correlations between the retrospective evaluation and the total experience of PA could be a consequence of the fact that the performance simply was not sufficient for real-world data.

Threats to external validity correspond to the generalisability of the experimental results. The preliminary results suggested that the modal emotion detected by acoustic analysis could be used as a predictor for the overall evaluation of an experience. However, this conclusion was drawn based on a sample of forty-one participants. Replications of this work with variation on the approaches taken for acoustic analysis as well as incorporating different experimental domains are needed to confirm the findings. This study was the first attempt to manually and automatically investigate emotional experiences from thinking aloud protocols.

4.6 Concluding Remarks

This chapter presented an exploratory study to assess emotional experiences based on the think aloud technique and automatic vocal analysis of the verbalised content. The employed methodologies were revolved around the theory of distinction between memory-based and moment-based approaches (Kahneman et al., 2003) for measuring experience. In this regard, it was proposed that emotional analysis of vocal expressions extracted from think aloud verbalisations is a moment-based approach, whereas applying retrospective questionnaire is a memory-based approach.

Previous research identified two dimensions, pleasure and arousal, as the characteristics of experience and in the construction of moment-based and memory-based assessments. In this study, in addition to the pleasure and arousal dimensions, the third dimension of the PAD model, dominance, we investigated as a construct of experience as well. The results showed that there was no correlation between memory-based and moment-based assessments of pleasure and arousal. However a positive correlation was found between the two measurement approaches for the dominance dimension. Hence, it was argued that the dominance dimension is relatively stable over time and therefore the memory and moment-based assessments are related.

In addition, the results showed that the retrospective evaluation of experience was correlated with the modal emotion obtained from the concurrent assessment of the entire interaction. Therefore, it can be inferred that the more positive the modal emotion is, the more positive the retrospective assessment of the experience will be. Overall, the main contribution of this exploratory study was to evaluate the effectiveness of deploying vocal analysis to assess people's emotional experiences when interacting with a system. This alternative approach can better utilise the traditional think aloud technique, benefiting usability and UX researchers and practitioners. However, more research needs to be carried out to substantiate and improve the results of this study.

Chapter 5

Study 2: Emotional Assessment of UX, Variation in Methods

Given the proven feasibility of deriving emotional responses from think aloud utterances with the use of automatic analysis approaches, the study presented in this chapter aimed to expand the extent of the applied methodology by taking different approaches on vocal analysis as well as tracking facial expressions. To verify the application of the proposed methodology, self-reports of emotions as well as vocal and facial expressions of 35 volunteers were captured during thinking aloud sessions and system interactions. A contrived e-commerce website selling laptops was developed as an environment to stimulate emotional and cognitive responses. Results of the analysis showed significant correlations between self-reports and measures of vocal expressions, where both methods were employed to assess users' emotional responses during their experiences. Moreover, classification of facial expressions to predict the self-report of emotions achieved at above 70% recognition accuracy for valence and dominance dimensions. Another aspect investigated was the effects of the presence or absence of task-related goals on the concurrent appraisal of an experience. Results suggested significant differences in the perception of emotional experiences between participants in the two groups of the presence or absence of task-related goals.

5.1 Aim of the Study

Results from the Study 1 (Chapter 4) suggested that emotional responses elicited during thinking aloud sessions and system interactions can be detected by using an automatic approach to vocal analysis. In that study, emotional responses were analysed automatically (based on the ML techniques on the patterns of acoustic features) and manually (human raters) through the recording of vocal expressions. However, the low recognition results demanded for further improvements in order to derive a more robust and reliable conclusion (Section 4.4.1). Additionally, due to the fact that audiovisual processing is superior to each single modal (Pantic and Rothkrantz, 2003), facial expressions were also collected and analysed. Multimodality becomes more beneficial where failure of one channel is recompensed by another channel and a message in one channel can be explained by that in another channel.

Assessing expressive responses, the primary research question for this study is to investigate to what extent automatic and concurrent measurements of those signals can reveal emotions and therefore their associations with the other components of UX. To address this question, facial expressions in addition to voice can be utilised to concurrently track the dynamic user emotional state in a usability test session (Zaman and Shrimpton-Smith, 2006). In this study, an inexpensive device, the Kinect sensor (Microsoft, 2015b), was used to collect the emotional reactions. Data collected from Kinect and the subjective measures were merged to create a preliminary dataset for predicting concurrent emotional responses.

Consequently, a second study was designed to expand the extent of the analysis with respect to the following research goals (RG):

- RG1: To improve the recognition accuracy of vocal expressions analysis by taking adaptive approaches (see Section 3.4.3);
- RG2: To employ facial expressions as an additional method for automatic assessment of emotions through classification techniques;
- RG3: To employ the discrete as well as the dimensional models of emotions for subjective and retrospective evaluation of emotional experiences.

Furthermore, as it was explained in section 2.3.2, research has emphasised the importance of adopting a combination of methods for measuring emotions. While each component of emotions can be assessed by a variety of methods, this study evaluated two components of emotions: 1) Subjective, using self-reported scales and 2) Motor expression by analysing vocal and facial expressions evoked during thinking aloud. In other words, the association between the subjective and expressive emotion components was investigated.

- RG4: To explore to what extent subjective and expressive components of CPM model (Scherer, 2005) reveal the same emotional states;

Finally, given the earlier empirical observations on situational aspect, it is expected that depending on the context, individuals form different emotional appraisal towards interactive experiences. Therefore:

- RG5: To investigate the influence of situation on emotional appraisal;

5.2 Experimental Design

5.2.1 Ethical Approval

Before a research study that involves human participants can take place, the ethics approval of the study is required to be issued through the University of Leicester ethics review system. Therefore a brief description of the study including the objectives, procedure and the type of data collection was submitted to the ethical committee. The process of reviewing and final approval took about three weeks. The reference for the approval of the ethics application is: 9876-ss887-computerscience.

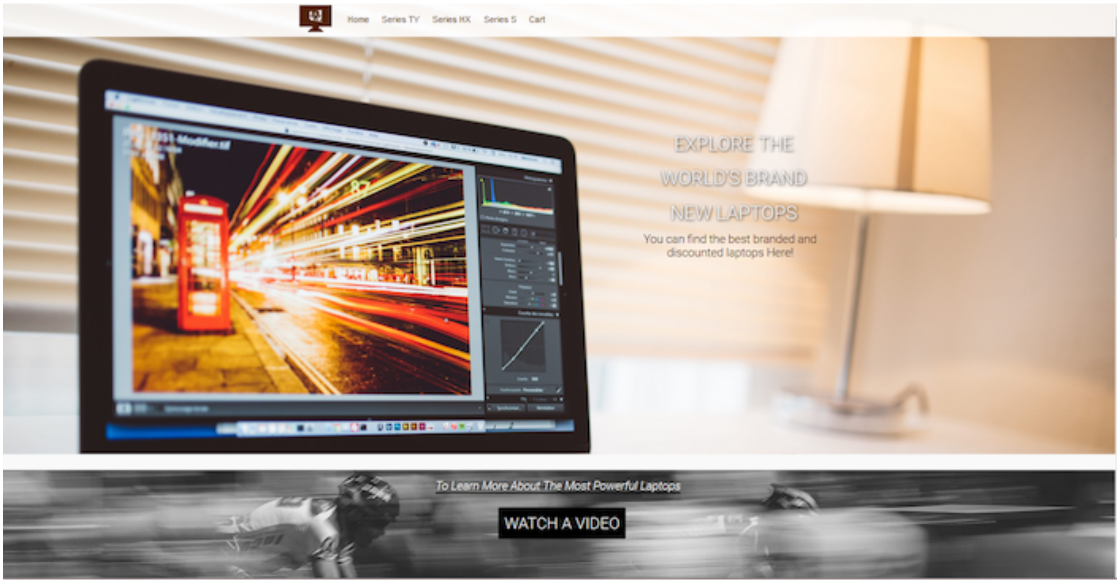
5.2.2 Participants

Thirty-five volunteers (female=16, male=19) participated. Invitation for participation were sent through personal contacts. Twenty two participants were between 18-25 and thirteen participants were between 25-35 years old. Twenty four of the participants were native English speakers and the rest were very fluent in English. The participants were students (undergraduate= 21 and postgraduate= 10) majoring in computer science, geography, law and management. Four participants were researchers working in geography, physics and engineering departments. Recruiting participants was carried out for a period of two and half months. Although data of all the participants were collected, data of five of them was discarded for the analysis due to failure in following the instructions.

5.2.3 Instruments: Software and Hardware

E-commerce Website

As it was discussed in the pilot study, the category to which a product belongs -hedonic or utilitarian - can account for consumers' attitude towards the product. Consequently, the ascription of a product in either of the two categories mediates the overall appraisal of the different UX constructs such as the perceived usability or aesthetics (Gross and Bongartz, 2012). On the other hand, there are products such as personal computer supplies that have been regarded with equally hedonic and utilitarian values (Crowley et al., 1992). Therefore, an e-commerce website selling



EXPLORE THE WORLD'S BRAND NEW LAPTOPS

You can find the best branded and discounted laptops Here!

To Learn More About The Most Powerful Laptops

Figure 5.1: Homepage of the CompDealers.com website.

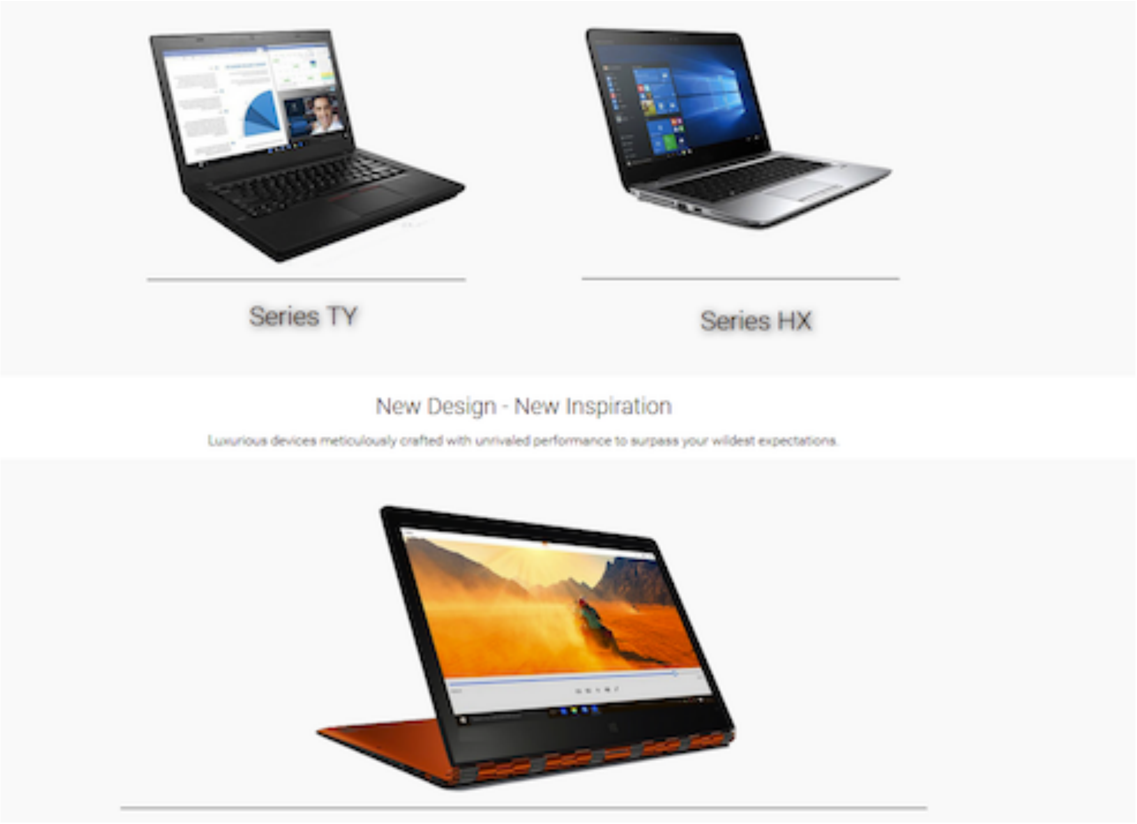
laptops (since they are more commonly used than personal computers in today's world) was used for this study.

I developed a contrived e-commerce website selling laptops, named as "CompDealers.com" to meet the purpose of this study. The website was created using the Wordpress platform (2002), which is an open-source online publishing platform and further customised by injecting additional HTML, javascript and CSS codes. The server of the CompDealers website was online for the duration of the study using the HostGator (2005), the web hosting platform.

The homepage of the website included a full-screen photo, a short video with a hedonic context to provide additional experiential information on the website's products (Figure 5.1). Three main categories of laptops were offered. Each laptop category was represented at the bottom of the homepage by an image and a link to a page with more detailed information (Figure 5.2).

Laptops were presented as generic as possible by removing any associated logos and text removed. A brand (integration of a name or a symbol) can induce identifiable products/services in consumers. Customers identify themselves with brands as a way for personality presence (Okazaki, 2006) and distinction from others. As a result, emotional and cognitive relationships can emerge between customers and brands (Gentile et al., 2007). These relations can impact on the evaluation of the system (Morgan-Thomas and Veloutsou, 2013). For example, the emotional ties between customers and brands could lead to satisfaction (Ha and Perks, 2005) and loyalty to a technology (Caruana and Ewing, 2010).

Consequently to eliminate the emotional and behavioural effect of brands, which could vary within individual customers (Gentile et al., 2007), dummy brands were



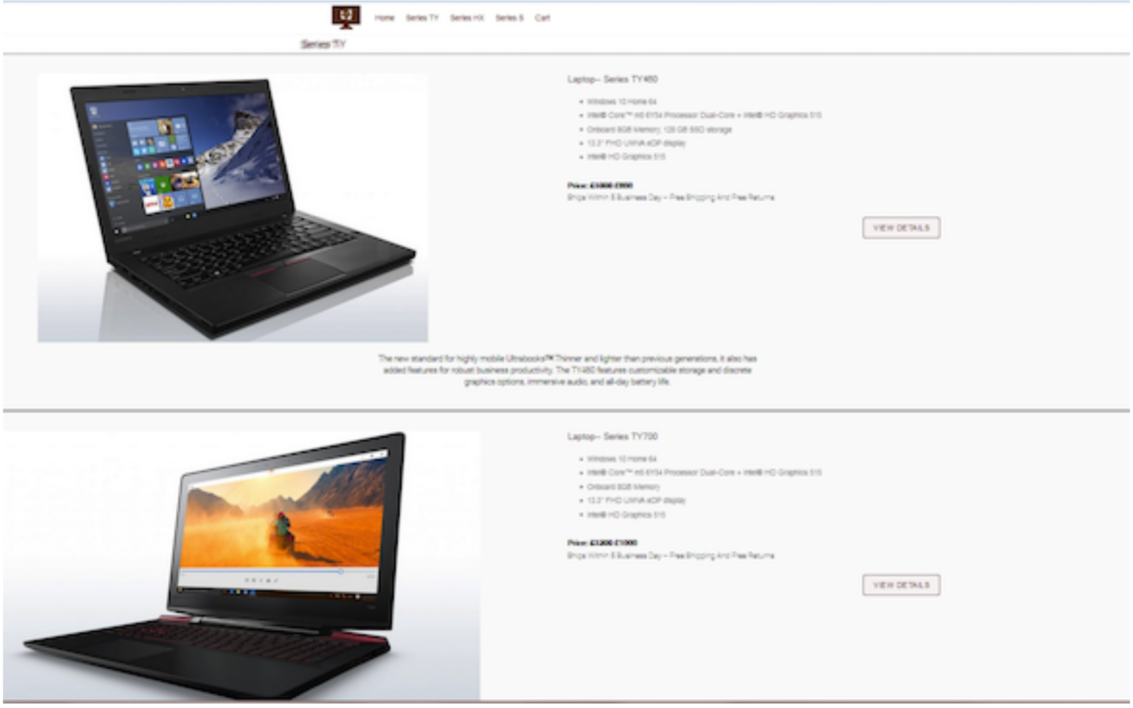
New Design - New Inspiration

Luxurious devices meticulously crafted with unrivaled performance to surpass your wildest expectations.

Figure 5.2: Three main categories of laptops on the homepage of the Compdealers.com website.

given to each category of laptops. With this approach, the overall attitude over a product that could be biased by a brand would be minimised (Crowley et al., 1992). Having fictitious brands was practiced in previous research as well to avoid confounding effects on participants' mental states (Sherwani and Stumpf, 2014).

All navigations for the first time from the homepage was followed by a prompting window asking participants to report their emotional status on the SAM scales.
Each laptop category had a webpage promoting two different versions of that category with two different prices (Figure 5.3). General specifications of the category were introduced in the page. However to see the more detailed specifications of each version, participants were required to navigate to another webpage (Figure 5.4 and 5.5).
When participants chose a laptop and were ready to complete the purchase, they clicked on the purchase button and reviewed the details of their selected product (Figure 5.6). The shopping process did not include a shopping form due to the lesson learned from the pilot study. In the pilot study, it was observed that most of the participants were not able to verbalise their states when filling out the shopping form. Some of them were only able to repeat the given mock credit card details.



Laptop– Series TY460

- Windows 10 Home 64
- Intel® Core™ m5 6Y54 Processor Dual-Core + Intel® HD Graphics 515
- Onboard 8GB Memory; 128 GB SSD storage
- 13.3" FHD UWVA eDP display
- Intel® HD Graphics 515

Price: £1000-£900
Ships Within 5 Business Day – Free Shipping And Free Returns

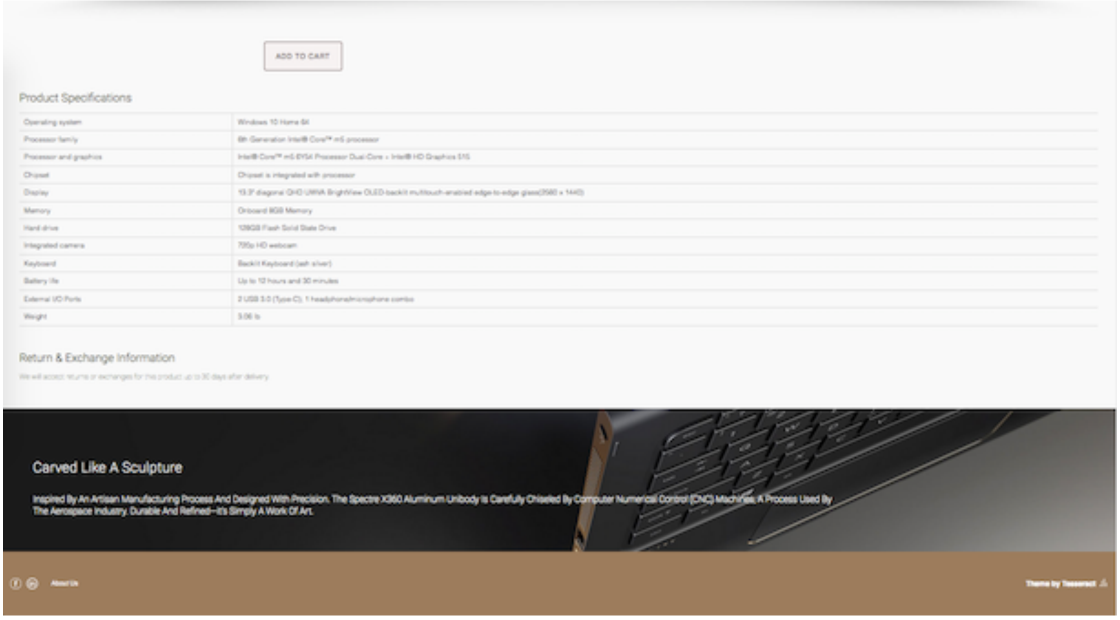
The new standard for highly mobile Ultrabooks™ Thinner and lighter than previous generations, it also has added features for robust business productivity. The TY460 features customizable storage and discrete graphics options, immersive audio, and all-day battery life.

Laptop– Series TY700

- Windows 10 Home 64
- Intel® Core™ m5 6Y54 Processor Dual-Core + Intel® HD Graphics 515
- Onboard 8GB Memory
- 13.3" FHD UWVA eDP display
- Intel® HD Graphics 515

Price: £1200-£1000
Ships Within 5 Business Day – Free Shipping And Free Returns

Figure 5.3: General specifications of a laptop category in the CompDealers.com website.



Product Specifications

Operating system	Windows 10 Home 64
Processor family	6th Generation Intel® Core™ m5 processor
Processor and graphics	Intel® Core™ m5 6Y54 Processor Dual-Core + Intel® HD Graphics 515
Chipset	Chipset is integrated with processor
Display	13.3" diagonal QHD UWVA BrightView OLED-backlit multitouch-enabled edge-to-edge glass(2560 x 1440)
Memory	Onboard 8GB Memory
Hard drive	128GB Flash Solid State Drive
Integrated camera	720p HD webcam
Keyboard	Backlit Keyboard (ash silver)
Battery life	Up to 12 hours and 30 minutes
External I/O Ports	2 USB 3.0 (Type C); 1 headphone/microphone combo
Weight	3.06 lb

Return & Exchange Information

We will accept returns or exchanges for this product up to 30 days after delivery

Carved Like A Sculpture

Inspired By An Artisan Manufacturing Process And Designed With Precision. The Spectre X360 Aluminum Unibody Is Carefully Chiseled By Computer Numerical Control (CNC) Machines, A Process Used By The Aerospace Industry. Durable And Refined—It's Simply A Work Of Art.

Figure 5.4: Detailed specifications of a laptop in the CompDealers.com website.

Figure 5.5: Slide show view of a laptop in the CompDealers.com website.

Therefore, for this experiment such a page with shopping form was not included as a part of the thinking aloud evaluation.

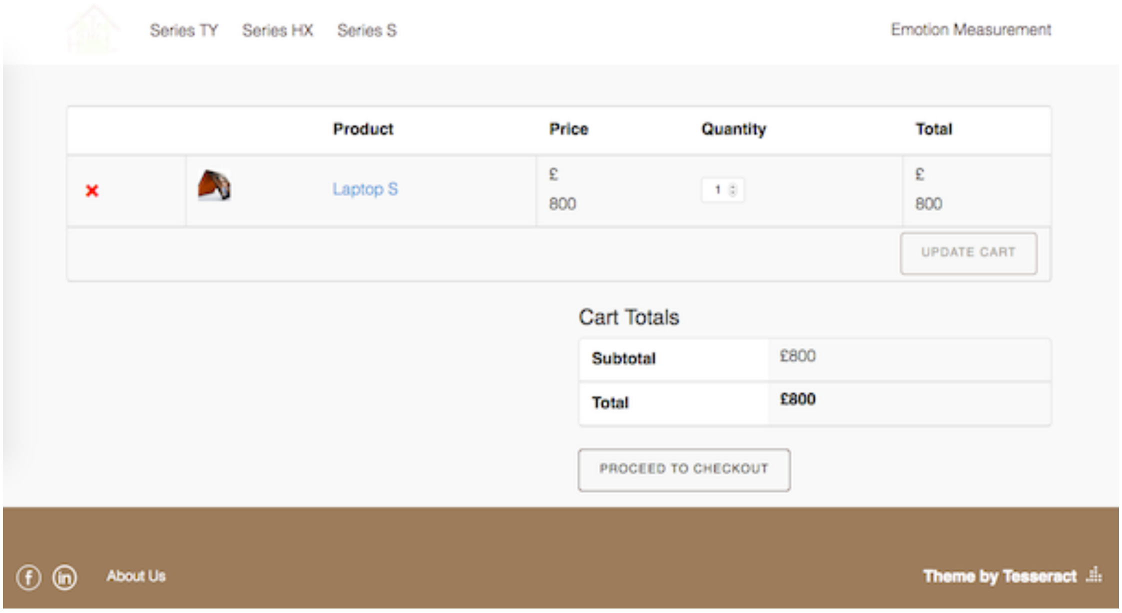
Logging Application

A WPF (Windows Presentation Foundation) desktop application was implemented using the Microsoft framework with C# as the programming language. This application served as an interface and background tool to execute a number of functionalities. As an interface to the Kinect sensor V2.0 (Microsoft, 2015b), which captures and streams facial data at a rate of 30 frames per second. Thus, a module was developed to extract and record data from Kinect. To develop this module, the Kinect Face Tracking Software Development Kit (SDK) was used to record the changes on the position of 17 facial movements, known as Animation Units (AUs). The outputted AUs from Kinect (Table 5.2) are a subset of Candid-3, the parameterised face model which has been used for coding of human faces (Ahlberg, 2001).

In this work, the changes on the positions of AUs from the neutral state were continuously logged within this module from the beginning of the interaction with the CompDealers.com website until the end of the experiment.

The second purpose of this application was to create a form in order to embed the SAM scales. This form would appear in the corner of the screen for participants to subjectively report the emotional states during their interaction.

Finally, an audio-recorder was integrated into the application for recording audio data with 44.1 kHz and 16 bits per sample format (This is the standard configuration



		Product	Price	Quantity	Total
×		Laptop S	£ 800	1	£ 800

Cart Totals

Subtotal	£800
Total	**£800**

Figure 5.6: View of the cart in the CompDealers.com website.

that has been used in studies for voice analysis such as the work by Ringeval and colleagues (2013)).

Hardware

The hardware included a Microsoft Kinect for Windows v2 sensor to record facial expressions. The Kinect sensor is connected to a Windows PC via a power supply and computer interface hub is dedicated to a USB 3.0 port. As the required operating system for Kinect is a Windows 8 or a later version, a desktop computer running Windows 10 was used. Audio data was captured by an external headset microphone.

5.2.4 Setting and Procedure

Participants were allocated and invited to individual lab-based sessions. The sessions were held in an office in the Department of Informatics, University of Leicester. A desktop computer was provided for participants to interact with the shopping website while wearing a headset microphone. The Kinect sensor was placed next to the computer, however it was adjusted for each individual to fully capture their facial expressions (Figure 5.7).

Each session began with welcoming participants and briefing about the objectives as well as the procedure of the study. In the introduction phase, the experimenter started with a presentation about the Kinect sensor and its purposes such as the type of data being collected by the sensor. Similar to the pilot study, the think aloud

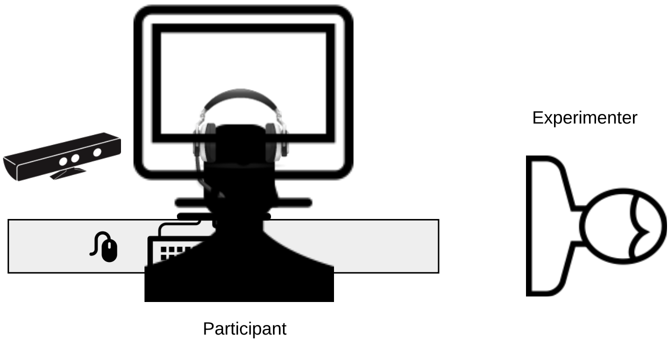


Figure 5.7: Experiment setup.

protocol was explained and practiced by a short trial session on the Amazon.com website.

Additionally, during the trial phase, participants were asked to fill the SAM questionnaire GUI, which was embedded in the corner of the screen. They were instructed to similarly complete the questionnaire at least two times during the actual experience:

- Anticipated peak: after navigating from the Home page of the website for the first time (mandatory)
- During: after visiting a laptop's specification page (optional)
- End: after navigating to the Cart when they want to place an order (mandatory)

Concurrent assessment of the subjective ratings was designed to be carried out at the transitions from one to another interface of the website. These transitions can provide “natural” moments for interruptions, therefore minimising influence on the current emotional state (Adamczyk and Bailey, 2004).

After participants indicated that they had understood all instructions, the consent form was given to be read and signed by participants. After signing the contest form, participants were asked to complete a questionnaire for background information, including items for evaluating the motivational orientation of participants based on a scale which will be explained in the subsequent sections.

Next, participants were asked to randomly read one of the following two scenarios and to think aloud while interacting with the CompDealers website. Note that immediately before and after reading the scenario, participants were asked to report their current emotional states using a paper-based version of the SAM scales.

Task-oriented scenario: “Imagine you have recently found a new job, which demands you to start immediately. However the company has informed you that they do not have laptops available for you to work with and so your project manager has asked you to find and order a laptop as soon as possible to start working right away. In this purchase there is no budget constraint but at the same time you are required to make sure the price is in a sensible range, therefore you should be able to justify your choice if you were asked to do so. In addition, you are strongly advised to carefully consider the required features and functionality of the laptop that you need for your work. This budget is being allocated once every two years to each employee and so you need to make sure you will consider the best possible combination of features that a laptop can have: such as performance, memory, weight and so on. In your field of work, you sometimes need to work long hours with your computer and you have experienced that the better a laptop performs, the better outcome you will achieve. In addition, you have in mind that sometimes you need to travel to different cities to meet and work with different clients. For you it is very important to be able to choose the best featured laptop with a reasonable price and more critically, to get the laptop as soon as possible in order to catch up with the other team members in this project."

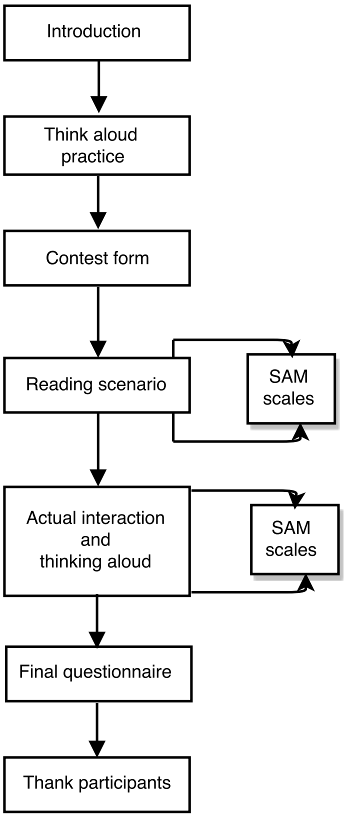


Figure 5.8: Procedure of the main study.

Fun-oriented scenario: "Imagine you have a good amount of savings. You have been planning to buy a laptop that you have been always desired. This laptop is going to be used for work as well as fun and entertainment. Today is the day that you feel like indulging yourself in a shopping experience, being relaxed and spend- ing time while browsing different models and designs of laptops and finally choose the one that you want. This is like a treat for you and all your hard work. You always enjoy browsing different products online especially for the things that you have been planning to have.
This laptop is going to be a perfect combination of functionally and look, in other words all the features that are in line with your desires. So you don't want to put any price constraint upon yourself as long as you know you are happy and going to enjoy your decision for this purchase."

The interaction phase involved participants to think aloud their thoughts and emotional experiences, as well as reporting their emotional states in the given GUI-based SAM scale.

Interactions were autonomously terminated by the participants, and next they were directed to complete a final questionnaire. The questionnaire was web-based service on the Qualtrics platform (2000) and included items for assessing emotions, usability, aesthetics and hedonic qualities of the experience.

At the end, all participants were thanked and compensated with a drink voucher in one of the University cafeterias. Each session took an hour on average.

5.2.5 Transcription and Segmentation

Thinking aloud data of all thirty participants was first transcribed and then segmented into units of analysis, as described before (Section 4.2.4): First, all the recorded audio files were chopped into chunks based on the silence pauses of longer than 1 second using the Audacity software. Next, two human raters were asked to separately evaluate each chunk as a unit (utterance), which should be meaningful and emotionally consistent throughout the chunk. If the chunk was evaluated as an utterance, the value of 1 was assigned, otherwise 0. Note that each rater had access to the transcripts and audio files.

One of the raters was an electrical engineer with having experience in the HCI field and the other one was myself. Hence the same raters as in the Study 1.

A number of discussion sessions was held to reach agreement on the segmented units or utterances. In this regard, the meaningless segments were discarded. The segments that were believed to consist of more than one utterance were split accordingly. Finally, the inter-rater agreement was computed using Cohen's Kappa= 0.74. In total 2309 segments were recognised as meaningful units for further analysis (mean=77, SD=45.8 utterance per participant). The average number of words per segment was 11.4 (SD= 2.4) and the average time of each segment was 4.4 seconds (SD= 0.49) across all participants.

5.3 Measurements

5.3.1 Emotion Assessment Methods

Since emotional experiences are the key constructs of UX, different approaches were employed in this study to assess such experiences. For retrospective assessment, subjective Self-reports were used, whereas for concurrent assessment, three methods

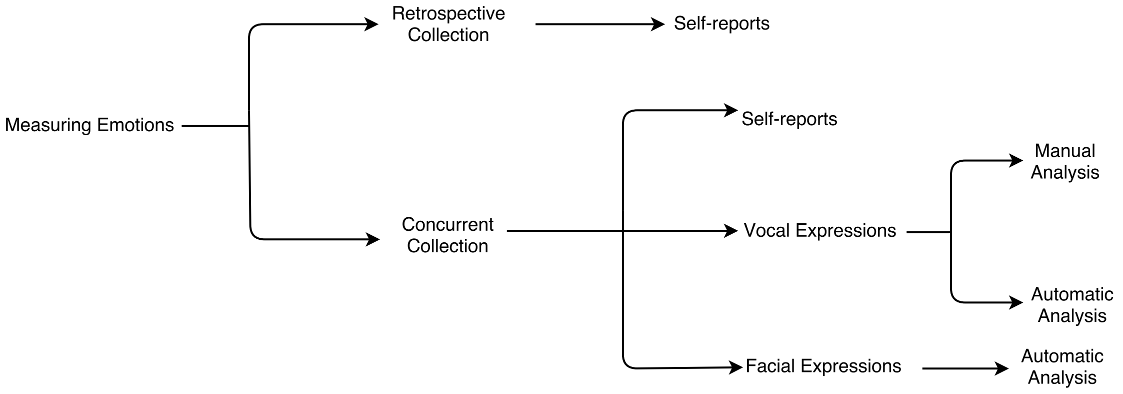


Figure 5.9: Diagram representing different assessment methods of emotions with respect to retrospective and concurrent approaches.

were employed, including subjective Self-reports; Vocal expressions using Manual human ratings and Automatic analysis; and Facial expressions by Automatic approaches. Collecting data from subjective ratings and motor expressions (voice and face) in the concurrent assessments examines the relationships between these two components of emotions. Figure 5.9 illustrates the categorisation of the methods.

Subjective Self-reports

In this study, SAM ratings were used as an account for the concurrent assessment of emotions based on the dimensional scheme during the actual interaction. To assess emotions retrospectively, the PAD sub-scale of Porat and Tractinsky questionnaire was employed. Note that, the Porat and Tractinsky scale has also items for measuring aesthetic and usability qualities of an experience, and as well as items for assessing the overall evaluation of a system, they will be discussed in Section 5.3.3.

The Differential Emotion Scale (DES) was another employed instrument for measuring emotions retrospectively, however within the discrete scheme (referred to section ??). To reduce the size of the questionnaire, the items describing disgust, contempt, fear, shame and guilt were removed for the emotional assessment of participants. The other emotions were selected according to the literature, where evidence of their occurrences was provided in the context of HCI. For instance, joy is an emotional experience that is usually evoked as a consequence of the appraisal of matching events towards related goals or as response to an aesthetically pleasing design (Desmet and Hekkert, 2007). On the other hand, anger, frustration and sadness are responses resulting from the appraisal of mismatching or hampering events (Desmet and Hekkert, 2007). Interest is associated with the desire to explore more, thus indicating an engaging experience and surprise is the desire to get inspired by or through the interaction (Olsson et al., 2013).

	1	2	3	4	5	6	7	8
1. Pleasure	0.93							
2. Arousal	0.74**	0.84						
3. Dominance	0.92**	0.70**	0.76					
4. Joy	0.73**	0.56**	0.69**	0.95				
5. Interest	0.50**	0.46*	0.54**	0.48**	0.64			
6. Surprise	0.40*	0.49**	0.44*	0.67**	0.37*	0.83		
7. Anger	-0.62**	-0.49**	-0.45*	-0.39*	ns	ns	0.85	
8. Sadness	-0.67**	ns	-0.54**	ns	ns	ns	0.68**	0.82

Table 5.1: Correlations between the constructs of PAD dimensions and DES emotion labels. Grey diagonal cells show Cronbach's alpha coefficient of reliability.

Internal consistency of each of the PAD dimensions was satisfactory: Cronbach's alpha of .93 for pleasure, .84 for arousal and .76 for dominance (the emotions labels in the Porat and Tractinsky scale were presented in Table 2.3). Based on this, the items of each dimension were averaged to represent a value for the dimension. Internal consistency for the chosen emotion categories from the DES scale was also calculated. Cronbach's alphas for all emotion labels were good (above .82), except for the interest category, Cronbach's alpha= .64, which is a moderate value.

Table 5.1 shows the correlation between the PAD sub-scales and the selected emotion labels from DES. As it can be seen from the table, for Pleasure, strong and positive correlations were found with joy and interest, whereas strong and negative correlations with anger and sadness. For Arousal, strong and positive correlations were confirmed with joy and surprise and a negative association with anger; and in terms of Dominance, strong and positive correlations were observed with joy and interest and a strong and negative relation with sadness. In other words, surprise was less associated with Pleasure; Sadness was not associated with Arousal and interest had a weaker correlation with Arousal; Finally surprise and anger had weaker associations with Dominance.

Furthermore, regarding the derived correlations among the emotion resulted from the subjective ratings, the not significantly correlated emotions were: sadness and joy; anger and interest; sadness and interest; surprise and anger; and surprise and sadness.

Note that that the both SAM and the Porat and Tractinsky's emotion sub-scale, are instruments for measuring emotions based on the PAD framework (Mehrabian, 1996). As compared to the PAD scale that consists of 18 items, SAM is a short and quick measurement instrument to efficiently assess responses to stimuli within an experimental session (Bradley and Lang, 1994).

Vocal Expressions Analysis

Two approaches were employed for the emotional assessment of vocal expressions: 1) Manual analysis through Human coding and 2) Automatic analysis.

Human coding: Two human raters (the same people from the segmentation task) individually coded the emotional value of thinking aloud data based on the PAD dimensions. However unlike Study 1 (Chapter 4), the coding scheme was followed on the basis of Pleasure (Negative/Positive), Arousal (Low/High) and Dominance (Low/Neutral/High).

Prior to the coding task, the two raters attended a meeting to gain a common understanding of the notion of each dimension of PAD. Having access to the transcripts and audio files, the raters were asked to evaluate the emotional value of each segment based on the binary Pleasure and Arousal and ternary Dominance dimensions. The following steps were undertaken:

For Participant$_{\mathrm{i}}$, whilst listening to the corresponding audio files:

1. Code the Pleasure of all segments in one go based on the Negative/Positive scheme
2. Code the Arousal of all segments in one go based on the Low/High scheme
3. Code the Dominance of all segments in one go based on the Low/Neutral/High scheme
4. Mark the neutral segments
5. Discuss the coding results with the other rater and recode, if necessary

Although the manual rating was a tedious and time consuming approach, the rationale was to establish ground truth data for the research questions related to measuring the vocal expressions. The Kappa value was computed to test the reliability of the coding between the two human raters. After a round of discussions, the Kappa was 0.73 for Pleasure, 0.80 for Arousal and 0.67 for Dominance.

Automatic analysis: The segmented data was analysed according to the methodology explained in the Section 3.4.3 of Chapter 3. Consequently, all segments were assessed with respect to the PAD dimensions in a binary manner for Pleasure (Negative/Positive) and Arousal (Low/High), whereas in a ternary (Low/Neutral/High) perspective for Dominance. The procedure for each dimension has been reviewed as follows:

Pleasure: An ensemble of SVMs was formed and the final outcome was determined via the majority weighted voting for binary decision Algorithm (3).

Arousal: With regard to Arousal, first the adaptive proposed approach based on concatenating the neutral references was employed Algorithm (3). Having identified neutral instances of each participant's verbalisations through the analysis of human coding, 10% of these instances were randomly taken out and concatenated to the IEMOCAP, the training corpus, for model creation. This method was proceeded in order to identify the Arousal samples with the label High or (1) with a greater probability (specificity). Next, for the remaining segments, ensemble of SVMs formed again based on the majority weighted voting for binary decision Algorithm (3) for determining the final label from the recognised results.

Dominance: Likewise, an ensemble of SVMs was built, however this time, a slightly different method was used to decide the recognised label from the multiple classifiers. For Dominance, in addition to having the binary labels Low (0) and High (1), the separate class N was allocated for the neutral samples. Hence the labels were determined as $c_i \in \{0, N, 1\}$. Similar to Pleasure and Arousal, weights were assigned to the outputs of each classifier, where the weight of the k classifier is denoted as: $w_k \in$ [0,1]. Running the majority weighted voting for ternary decision Algorithm (4), the final decision class $c_i \in \{0, N, 1\}$ was determined for this dimension.

Recognition performance: Manual human ratings were used as a benchmark against which the automatic recognition were validated. Similar to Study 1 (Section 4.4.1), the overall accuracy for each of the three dimensions Pleasure, Arousal and Dominance was calculated. As results, the obtained accuracy for each of the Pleasure, Arousal and Dominance was 0.66, 0.70 and 0.41 respectively. It should be noted that classifications on Pleasure and Arousal were binary tasks and on Dominance was a ternary class-problem task.

Facial Expressions Analysis

The Microsoft Kinect sensor V2.0 (Microsoft, 2015b) is capable of capturing human faces along with bodily features in real-time. This sensor comes with a robust SDK that provides Application Program Interface (API) to track facial traits in terms of Face Shape Animations (FSA). FSA data tracked by Kinect constitutes a subset of the AUs from the original FACS. The facial features and their corresponding AUs

Attribute	Features	Kinect AUs
1	Jaw Lowerer	AU0
2	Lip Pucker	AU1
3	Jaw Slide	AU2
4	Lip Stretcher (Left, Right)	AU3, AU4
5	Lip Corner Puller (Left, Right)	AU5, AU6
6	Lip Corner Depressor (Left, Right)	AU7, AU8
7	Cheek Puff (Left, Right)	AU9, AU10
8	Eye Closed (Left, Right)	AU11, AU12
9	Brow Lowerer (Left, Right)	AU13, AU14
10	Lower Lip Depressor(Left, Right)	AU15, AU16

Table 5.2: Microsoft Kinect v2 face shape animation units (AUs) (Microsoft, 2015a). The attribute column indicates the assigned number of each attribute in the created ARFF file.

from Kinect V2.0 are presented in Table 5.2. Each AU is expressed as a numeric weight varying between -1 and +1, and the neutral states of AUs are normally assigned 0.

As it was explained before (Section 5.2.3, Logging Application), video sequences can be acquired approximately as 30 frames per second using the Kinect sensor. On each frame, face detection and feature extraction are performed by the Kinect face tracking SDK. The outputs of the face tracking are AUs, which could be used for inferring facial expressions. Therefore, collecting and analysing output data from the Kinect sensor was another methodology used for measuring emotions.

As it can be seen from Table 5.2, Kinect's output differentiates between the right and left sides of the face (e.g. left and right side of the Lip Stretcher). However, the original FACS did not distinguish between the left and right side of the facial movements described by the corresponding AUs (e.g a single AU describing the Lip Stretcher) to infer emotions (Ekman and Friesen, 1977). Previous work such as (Mehu and Scherer, 2015) also used a single value describing each AU related to the both left and right side of the face. Hence, in this work, the average value obtained from the left and right AUs was used as well.

Previous work has shown correlation between facial expressions and self-reported emotions (Ekman and Rosenberg, 1997). Mehu and Scherer (2015) demonstrated the associations between PAD dimensions and AUs, where AUs were coded by certified human coders of the FACS. Relying on these theoretical research, three training sets were created using the AUs and the self-reported SAM ratings as ground truths for Pleasure, Arousal and Dominance dimensions. To extract the facial expressions coincidental with the self-reports, the time-stamp of the self-reports with the facial data was checked. Hence, the facial features elicited in a specific time-stamp were associated with the data of self-reports.

All of the AU features were formalised as an attribute-relation file format (ARFF) for WEKA. Next, an emotion class (such as a class describing pleasure, arousal or dominance) attribute, was assigned to each ARFF file indicating the emotional label of that file. As results, a training dataset consisting of 111 instances was constructed (*Note* that for each of the PAD dimensions, a separate dataset was created). Similar to the vocal analysis, the emotional labels for Pleasure were Negative/Positive, for Arousal were Low/High and for dominance were Low/Neutral/High.

Two different supervised machine learning algorithms were used to construct the classifiers. The classifiers were K-Nearest Neighbours (KNN) and SVM with SMO training algorithm. KNN is a simple classification algorithm that stores all available cases and classifies new cases based on the majority vote of the K nearest neighbours. To choose an optimal value for K in KNN and C in SVM classifiers, a range of values for both the parameters were tested within a grid-search mechanism.

Finally, in addition to the AUs, Kinect outputs some information about the state (happy and engaged) or appearance (such as wearing glasses, mouth open, and the left or the right eye closed) of a user's face. The corresponding value of each stance (happy or engaged as state and wearing glasses as appearance) estimated by Kinect varies from yes (certain true property), no (certain false property), maybe (most likely the property is true) and unknown (Microsoft, 2015a). Consequently, the happy and engaged states outputted from Kinect could be used for concurrently assessing the PAD dimensions.

Hypothesis:

The discrepancy between actual experienced emotions and the retrospective reports of them has been debated in a number of studies (as it was discussed in Section 5.1). The results of Study 1 (Chapter 4) also suggested that the concurrent assessment of emotions using vocal expressions do not correlate to the corresponding retrospective self-reports. Given the evidence, it is expected that the same differences between Retrospective and Concurrent assessments are observed when subjective ratings, SAM, are employed for measuring emotions during system interaction. Therefore:

- H1. Concurrent and retrospective self-reported assessments of the PAD dimensions (Pleasure, H1a; Arousal, H1b; Dominance, H1c), denoted as $\text{PAD}_{\text{Self-Con}}$ and $\text{PAD}_{\text{Self-Ret}}$, do not correlate significantly.

On the other hand, it is hypothesised that two different concurrent assessment methods of emotions show significantly positive correlation. Hence, regarding to the concurrent assessment of emotions by vocal analysis, two hypotheses are formulated as follows:

- H2.1. Concurrent self-reports and Manual analysis of vocal expressions of the PAD dimensions (Pleasure, H2.1a; Arousal, H2.1b; Dominance, H2.1c), denoted as $\text{PAD}_{\text{Self-Con}}$ and $\text{PAD}_{\text{Voc-Man}}$), correlate positively and significantly.
- H2.2. Concurrent self-reports and Automatic analysis of vocal expressions of the PAD dimensions (Pleasure, H2.2a; Arousal, H2.2b; Dominance, H2.2c), denoted as $\text{PAD}_{\text{Self-Con}}$ and $\text{PAD}_{\text{Voc-Ac}}$), correlate positively and significantly.

In addition, using the automatic analysis of vocal expressions, it is hypothesised that the Automatic and Manual ratings should associate:

- H3. Manual and Automatic analysis of vocal expressions of the PAD dimensions (Pleasure, H3a; Arousal, H3b; Dominance, H3c), denoted as $\text{PAD}_{\text{Voc-Man}}$ and $\text{PAD}_{\text{Voc-Ac}}$, correlate positively and significantly.

Similar to the hypothesis H1, it is expected that the Manual ratings of emotions do not correlate significantly with the corresponding Retrospective self-reports:

- H4. Retrospective self-reports and Manual analysis of vocal expressions of the PAD dimensions (Pleasure, H4a; Arousal, H4b; Dominance, H4c), denoted as $\text{PAD}_{\text{Self-Ret}}$ and $\text{PAD}_{\text{Voc-Man}}$, do not correlate significantly.
- H5. Retrospective self-reports and Automatic analysis of vocal expressions of the PAD dimensions (Pleasure, H5a; Arousal, H5b; Dominance, H5c), denoted as $\text{PAD}_{\text{Self-Ret}}$ and $\text{PAD}_{\text{Voc-Ac}}$, do not correlate significantly.

Finally, using the Kinect sensor, it is predicted that the Kinect's output (frames of happy) correlates significantly to the concurrent self-reports measures. Therefore, it is hypothesised that:

- H6. The number of happy frames outputted from Kinect is positively and significantly correlated with the value for Pleasure, H6a; Arousal, H6b; and Dominance, H6c.

It should be noted that aligned with the previous study (4), the concept of Total experience (Algorithm 5) was used to represent the overall emotional experiences derived from the different concurrent methods.

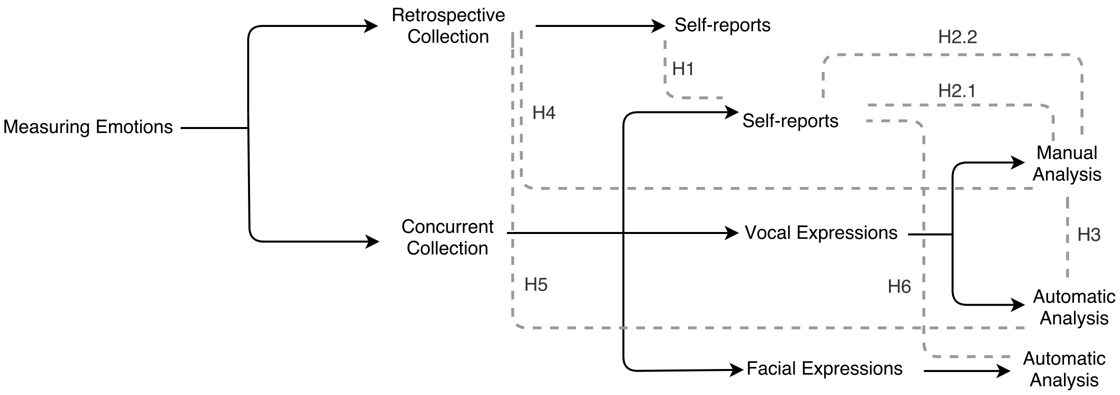


Figure 5.10: Diagram representing hypotheses of this study formulated based on different assessment methods.

		Concurrent Self–reports		
		Pleasure	Arousal	Dominance
Retrospective Self–reports	Pleasure	0.45*	0.29	0.55**
	Arousal	0.66**	0.55**	0.65**
	Dominance	0.31	0.30	0.40*

Table 5.3: Spearman's rank-order correlations table of Retrospective and Concurrent self–reported ratings of the PAD dimensions. **. Correlation is significant at the 0.01 level and *. Correlation is significant at the 0.05 level (2-tailed).

Results:

To examine the associations between the measures of the PAD dimensions assessed by different approaches, a number of Spearman's rank-order correlations were performed.

H1: Correlations between $\mathrm{PAD}_{\text{Self-Con}}$ and $\mathrm{PAD}_{\text{Self-Ret}}$ were computed (Table 5.3). Results showed significant and positive correlations between Retrospective and Concurrent ratings in measuring Pleasure (r=0.45, $p < 0.05$), Arousal (r=0.55, $p < 0.05$) and Dominance (r=0.40, $p < 0.05$). Consequently, H1a, H1b and H1c were rejected. In addition to the existing correlations between the corresponding $\mathrm{PAD}_{\text{Self-Con}}$ and $\mathrm{PAD}_{\text{Self-Ret}}$, strong correlations were observed between $\mathrm{A}_{\text{Self-Ret}}$ and $\mathrm{P}_{\text{Self-Con}}$, (r=0.66, $p < 0.01$) and between $\mathrm{A}_{\text{Self-Ret}}$ and $\mathrm{D}_{\text{Self-Con}}$, (r=0.59, $p < 0.01$) (See Table 5.3). Overall, the results were contradicting those of the previous work (Bruun and Ahm, 2015), where no significant correlations had been confirmed between Concurrent and Retrospective ratings.

Moreover, the mean of Concurrent and Retrospective ratings were compared using the Wilcoxon signed-rank test. No significant difference was observed in measuring Pleasure between Retrospective (Mean=5.97, SD=1.70) and Concurrent ratings (Mean=6.40, SD=1.58). In terms of Arousal, the Retrospective rating (Mean=5.36, SD=1.25) was significantly lower than the Concurrent measure (Mean=6.11, SD=1.25). This difference was observed by: Z = -2.83, $p < 0.001$. Similar to Pleasure, no significant difference was observed in measuring Dominance between Retrospective (Mean=5.87, SD=1.26) and Concurrent ratings (Mean=6.08, SD=1.44).

H2.1: To determine whether $\text{PAD}_{\text{Self-Con}}$ correspond to $\text{PAD}_{\text{Voc-Man}}$, correlations between these methods on the PAD dimensions were calculated. Results showed no significant associations between the Manual ratings of vocal expressions and the Concurrent self-reports on assessing: Pleasure (r=0.34, p=0.078), Arousal (r=0.22, p=0.25) and Dominance (r=0.19, p=0.42). Consequently, the H2.1 was rejected.

However, having a further investigation in individual data showed that there is an unusual difference between the measures of $\text{A}_{\text{Self-Con}}$ and the measures of $\text{A}_{\text{Voc-Man}}$ of four participants. In this regard, the average of the differences (difference= $\text{A}_{\text{Voc-Con}}$ - $\text{A}_{\text{Voc-Man}}$) of the four participants was 0.5 (SD=0.02), whereas the average of the same differences for the rest of participants (n=26) was 0.21 (SD=0.14). Therefore, the data of these participants was removed (outlier data) and the correlation was again computed. This time, the results showed a significant correlation between $\text{A}_{\text{Voc-Con}}$ and $\text{A}_{\text{Voc-Man}}$ (r=0.40, p=0.045).
The four participants were Computer Science students and had previously taken part in study 1. The notable differences in their arousal measures could be traced back to the plausible effect of social desirability, wherein participants tend to act in a particular way to achieve what researcher/experimenter wants rather than their actual responses (Phillips and Clancy, 1972). In other words, some participants make some assumption of the study's hypotheses, which may not correspond to the actual ones, and have the tendency to respond accordingly. In particular, the effects of social desirability is expected to distort the responses of self-reported surveys (Phillips and Clancy, 1972).

With regard to H2.1a and H2.1c, similar to arousal, the differences between the measures of $\text{P}_{\text{Self-Con}}$ and the measure $\text{P}_{\text{Voc-Man}}$, and the measure of $\text{D}_{\text{Self-Con}}$ and the measure $\text{D}_{\text{Voc-Man}}$ were calculated. However, no such discrepancies were found among the data of these four participants for these two dimensions. Although the p-value for H2.1a became close to significant (r=0.35, p=0.077)
Figure 5.11, the first row, shows the correlations between $\text{PA}_{\text{Self-Con}}$ and $\text{PA}_{\text{Voc-Man}}$. *Note* that the Concurrent self-reports were collected by using SAM in scales from 1 to 9. In order to better illustrate the correlations in Figure 5.11, Pleasure and Arousal

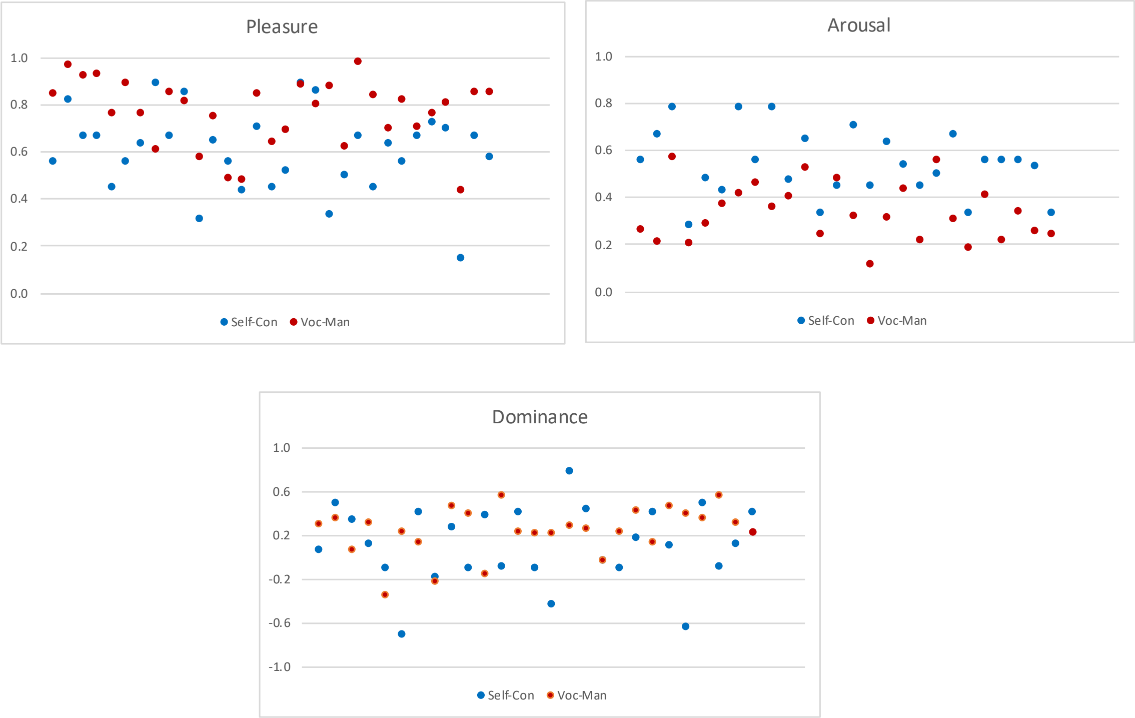


Figure 5.11: Spearman's rank-order correlations obtained between the measures of Pleasure (left), Arousal (right) and Dominance (bottom) assessed by Concurrent self-reports (Self-Con) and Manual ratings of vocal expressions (Voc-Man). Pleasure and Arousal have been demonstrated in the scale of 0 (Negative/Low) to 1 (Positive/High) and Dominance in the scale of -1(Low) to 0 (Neutral) to 1 (High).

ratings of SAMs were rescaled to the range of 0 to 1, similar to the corresponding measures in the Manual ratings. Accordingly, the relationship in the measures of Dominance by $D_{\text{Self-Con}}$ and $D_{\text{Voc-Man}}$ has been illustrated in the second row of Figure 5.11. The SAM ratings of dominance were rescaled to the range of -1 to 1 in order to correspond to their counterparts in Manual ratings of vocal expressions (where -1 was assigned to Low dominance, 0 to Neutral and +1 to High dominance).

In addition, the mean and standard deviation of the $PAD_{\text{Self-Con}}$ and $PAD_{\text{Voc-Man}}$ were computed (See Table 5.5). While self-reports and manual ratings are both concurrent assessments, across all participants, Pleasure and Dominance were respectively measured as positive and high by the both methods, and Arousal as high by self-reports and as low by manual ratings (Table 5.5).

With regard to the significant differences, $P_{\text{Voc-Man}}$ was significantly greater than $P_{\text{Self-Con}}$: $Z = 4.46$, $p < 0.001$; $A_{\text{Voc-Man}}$ was significantly lower than $A_{\text{Self-Con}}$: $Z = -4.45$, $p < 0.001$; and $D_{\text{Voc-Man}}$ was significantly lower than $D_{\text{Self-Con}}$: $Z = -4.28$, $p < 0.001$.

	$P_{\text{Self-Con}}$	$P_{\text{Voc-Man}}$	$A_{\text{Self-Con}}$	$A_{\text{Voc-Man}}$	$D_{\text{Self-Con}}$	$D_{\text{Voc-Man}}$
Mean	6.40	0.77	6.11	0.34	6.08	0.59
SD	1.58	0.16	1.25	0.12	1.44	0.12

Table 5.4: The mean and standard deviation of the $PAD_{\text{Self-Con}}$ in 1 to 9 scales (and 0 to 1); and the $PAD_{\text{Voc-Man}}$ in 0 to 1 scales.

	$P_{\text{Self-Con}}$	$P_{\text{Voc-Ac}}$	$A_{\text{Self-Con}}$	$A_{\text{Voc-Ac}}$	$D_{\text{Self-Con}}$	$D_{\text{Voc-Ac}}$
Mean	6.40	0.47	6.11	0.55	6.08	0.48
SD	1.58	0.11	1.25	0.22	1.44	0.12

Table 5.5: The mean and standard deviation of the $PAD_{\text{Self-Con}}$ in 1 to 9 scales (and 0 to 1); and the $PAD_{\text{Voc-Ac}}$ in 0 to 1 scales.

H2.2: Concerning the associations between $PAD_{\text{Self-Con}}$ and $PAD_{\text{Voc-Ac}}$, the only significant correlation was confirmed in measuring Pleasure (r=0.40, p=0.04), hence H2.2a was accepted and H2.2b and H2.2c were rejected. With regard to the overall value of the PAD, as opposed to the Manual analysis, Pleasure and Dominance have been assessed as negative and low, whereas Arousal as high by the Automatic analysis.

The positive association between $P_{\text{Self-Con}}$ and $P_{\text{Voc-Ac}}$ can be used as evidence for proposing relationships between the subjective and expressive components in Scherer's CPM model.

H3: Correlations between the Manual and Automatic assessment of vocal expressions of the PAD dimensions were examined. Significant correlations were confirmed between $P_{\text{Voc-Man}}$ and $P_{\text{Voc-Ac}}$ (r=0.42, p=0.034) and between $A_{\text{Voc-Man}}$ and $A_{\text{Voc-Ac}}$ (r=0.60, p=0.002). However no significant correlation was found between $D_{\text{Voc-Man}}$ and $D_{\text{Voc-Ac}}$ (r=0.26, p=0.24). Consequently, H4a and H4b were confirmed, and H4c was rejected. These correlations have been illustrated in Figure 5.12).

H4: While hypothesis H1 examined the associations between Concurrent and Retrospective assessments by using self-reports, in H4, the associations between $PAD_{\text{Voc-Man}}$ and $PAD_{\text{Self-Ret}}$ were investigated (Table 5.6).
From Table 5.6, it can be observed that correlations exist for Pleasure and Dominance dimensions, when they are measured 1) by the Manual assessments of vocal expression and by 2) Retrospective self-reports. Similar associations were found for these two dimensions, where self-reports were used as the concurrent assessment

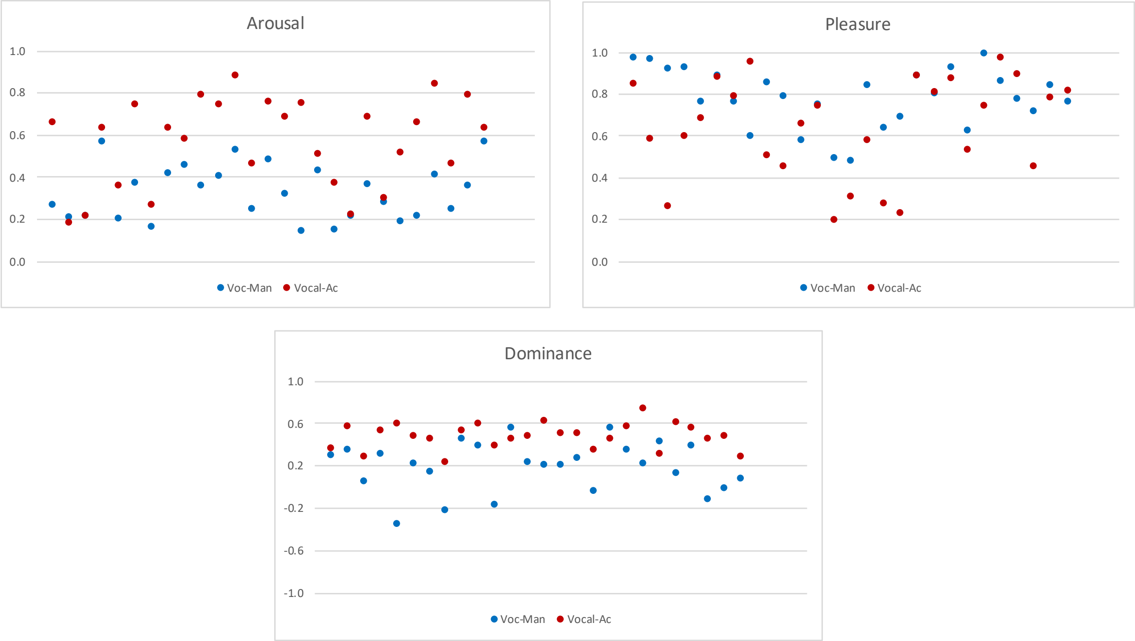


Figure 5.12: Spearman's rank-order correlations obtained between the measures of Pleasure (left), Arousal (right) and Dominance (bottom) assessed by Manual ratings (Voc-Man) and Automatic analysis (Voc-Ac) of vocal expressions. Pleasure and Arousal have been demonstrated in the scale of 0 (Negative/ Low) to 1 (Positive/ High) and Dominance in the scale of -1(Low) to 0 (Neutral) to 1 (High).

		Manual Assessments of Vocal Expressions		
		Pleasure	Arousal	Dominance
Retrospective Self–reports	Pleasure	0.49**	-0.07	0.39
	Arousal	0.48*	-0.14	0.34
	Dominance	0.48*	-0.09	0.46*

Table 5.6: Spearman's rank-order correlations table of Retrospective self-reports and Manual assessment of vocal expressions based on the PAD dimensions. **. Correlation is significant at the 0.01 level and *. Correlation is significant at the 0.05 level (2-tailed).

Training	Pleasure	Arousal	Dominance
SVM	68.5	63.9	67.6
KNN	74.8	66.7	78.4
Mean	71.7	65.3	73.0

Table 5.7: Accuracy of the SVM and KNN classification models, trained based on the concurrent PAD. 10-fold Stratified classification.

methods (Table 5.3). Moreover, Retrospective assessment of Arousal and Dominance were correlated with the Manual assessment of Pleasure by vocal expressions.

H5: Furthermore, correlations between Retrospective assessments, $\text{PAD}_{\text{Self-Ret}}$ and Automatic analysis of vocal expressions, $\text{PAD}_{\text{Voc-Ac}}$ were examined, however no strong associations were observed in measuring Pleasure, Arousal or Dominance. Hence, H5a, H5b and H5c were accepted accordingly.
Overall, the results imply that while Concurrent and Retrospective assessments of the PAD by using self-reports correlate, these correlations do not always hold, when self-report ratings are substituted by other methods, such as vocal analysis. This finding could imply that participants could remember what they had rated before. However, they could not remember how they voiced.

H6: As it was explained earlier, the Kinect sensor outputs a state of happy, when it predicts with certainty a participant is happy for a duration of each capturing frame. Similar to vocal analysis, a normalised value was calculated to represent the overall happy state for each participant. Then the correlations between the Kinect output and the self-reports were computed, given the happy state as positive Pleasure, high Arousal and high Dominance. However, no correlations were confirmed in measuring Pleasure (r=0.12), Arousal (r=0.16) and Dominance (r=0.13). Hence, H6a, H6b and H6c were rejected.

Classification on Facial Expressions: SVM and KNN algorithms were separately trained for each of PAD dimensions based on the corresponding concurrent self-reports and the AU attributes in a stratified 10-fold cross validation paradigm. Table 5.7 summarises the accuracy of the classifications performances. As it can be observed from results, the KNN classifier outperformed SVM for the three dimensions Pleasure, Arousal and Dominance. The better performance of KNN over SVM can be related to the size of the dataset, which was built out of 111 instances. In previous research it was observed that KNN algorithms usually outperform SVM, where the size of the dataset is relatively small (e.g. for the sizes between 50-200 instances) (Raikwal and Saxena, 2012).

Training	Pleasure	Arousal	Dominance
SVM	62.1	62.1	68.9
KNN	72.4	65.5	75.8
Mean	67.3	63.8	72.3

Table 5.8: Prediction accuracy on the retrospective PAD using SVM and KNN classifiers generated from the concurrent PAD.

This experiment's results suggested an average accuracy of 71.7% for Pleasure, 65.3% for Arousal and 73% for Dominance. Furthermore, Dominance and Pleasure classifications showed better accuracy performances than Arousal based on the data obtained from facial expressions. This pattern could be observed from the results of the both KNN and SVM classifiers.

Additionally, the classification models created from the facial expressions elicited during interaction and the concurrent PAD ratings (Train) were used for prediction on the retrospective ratings of PAD (Test). Results of the predictions have been illustrated in Table 5.8. In this regard, the models were able to accurately predict retrospective PAD up to 72.4%, 65.5% and 72.3% for each of the Pleasure, Arousal and Dominance respectively. As suggested from the results, the prediction accuracies obtained for the test dataset are in same ranges as the classification accuracies of the models developed from the training set. This outcome could suggest that facial expressions described by AUs and elicited during the interaction could be used for the prediction of the retrospective PAD self-reports.

Regression Analysis: Linear regressions were conducted to determine which particular AUs contribute most to predict each of the Pleasure, Arousal and Dominance from the facial expressions. Analysis showed that Jaw Slide, Jaw Lowerer and Lip Stretcher significantly contributed for predicting Pleasure ($R^2 = 0.27$, $F(110,100) = 3.56$, $p < 0.001$). Previous research had also suggested a positive correlation between lip stretching and pleasure (Frijda and Philipszoon, 1963). For Arousal, Lip Corner Depressor and Brow Lowerer were the significant contributors ($R^2 = 0.29$, $F(110,100) = 3.98$, $p < 0.001$). A negative and significant association was found between Arousal and Brow Lowerer with Arousal, which is in line with prior research, where it was shown that brow lowering indicates a high arousal state such as a thoughtful state (Littlewort et al., 2011) or a negative emotion like confusion (D'Mello et al., 2014) or frustration (Grafsgaard et al., 2013). Finally, for Dominance, Jaw Lowerer and Lip Stretcher were the most important AUs contributing in the prediction model ($R^2 = 0.27$, $F(110,100) = 3.73$, $p < 0.001$).

5.3.2 Products' Qualities

As it was explained earlier in this thesis (section 2.1.1), some basic features such as colour, music, and graphic contribute to fun and enjoyment aspects of an experience, which in turn influences the overall satisfaction.

In this study, Attrakdiff2 consisting of ten items (four items for each hedonic and pragmatic quality, two items for beauty and goodness (Hassenzahl and Monk, 2010) and the Porat and Tractinsky (which includes classic and expressive aesthetics subcategories) scale were used for measuring the perceived usability, hedonic and aesthetics qualities. Internal consistency (Cronbach's Alpha) within the items of each sub-category of the employed scales were computed. Cronbach's Alpha for the hedonic and pragmatic quality scales was good (above 0.80), as well as for the expressive aesthetics, usability and attitude scales (above 0.80). The value of Alpha was rather questionable for the classic aesthetics (0.64). However, by removing the item "Symmetrical" from this sub-scale, the internal consistency became good (0.86). As a result, a mean score was calculated from the items of each sub-scale of Attrakdiff2 and of the Porat and Tractinsky questionnaire.

To examine to what degree a product's perceived qualities are predictable from the emotional measures of the concurrent methods, a number of correlations were calculated. Table 5.9 shows the correlations between the concurrent and subjective self-reports of emotions based on SAM and the product's perceived qualities based on the Attrakdiff2 and Porat and Tractinsky scales. Results showed significant correlations between hedonic quality and Arousal (r=0.46, $p < 0.01$), and hedonic quality and Dominance (r=0.41, $p < 0.05$). This suggests that higher levels of hedonic quality leads to higher level of perceived Arousal. However, pragmatic quality did not have any effects on emotional experiences, whereas, usability was strongly correlated with Pleasure, Arousal and Dominance dimensions.

The inconsistent correlations between the dimensional emotions and pragmatic and usability of the system imply that pragmatic quality of Attrakdiff2 and usability of Porat and Tractinsky scales do not measure exactly the same quality. While the usability quality of the Porat and Tractinsky scale emphasises on the ease of use and navigation of a system, the pragmatic quality of Attrakdiff2 focuses on the capability of a product/service to achieve the related goals of system usage. In this study, the ease of use and navigation of the system resulted in higher levels of perceived pleasure (r=0.49, $p < 0.01$) and dominance (r=0.60, $p < 0.01$), whereas goal achievement did not impact on the emotional experiences particularly.

Furthermore, perceived beauty was strongly associated with arousal. Comparably, correlations were observed between arousal and both classic and expressive aesthet-

AttrakDiff2	Pleasure	Arousal	Dominance
Pragmatic	0.32	0.03	0.35
Hedonic	0.32	0.46**	0.41*
Beauty	0.13	0.38*	0.26
Goodness	0.52**	0.35	0.58**
Porat & Tractinsky	Pleasure	Arousal	Dominance
Usability	0.49**	0.37*	0.60**
Classic Aes	0.27	0.50**	0.02
Expressive Aes	0.56**	0.59**	0.64**
Attitude	0.53**	0.33	0.65**

Table 5.9: Spearman's rank-order correlations between concurrent and subjective Self-reports of emotions using SAM and the retrospective product's (CompDealeres.com) perceived qualities based on the Attrakdiff2 and Porat and Tractinsky scales. **. Correlation is significant at the 0.01 level and *. Correlation is significant at the 0.05 level.

ics. Although, expressive aesthetics was significantly associated with pleasure and dominance as well.

Goodness measured by AttrakDiff2 was significantly correlated with pleasure and dominance. A similar trend of relations was observed between pleasure-dominance dimensions and attitude assessed by Porat and Tractinsky's scale. The significant association between pleasure and attitude was consistent with the findings of previous work (Porat and Tractinsky, 2012). In this regard, higher levels of perceived pleasure increased participants' approach response about the system. Likewise, higher levels of perceived dominance had the same effect on participants' overall judgment about the system.

Table 5.10 shows the correlations between the emotions derived from concurrent vocal expressions with the Manual human coding approach and the product's perceived qualities. Similar to the use of subjective self-reports of emotions, correlations can be seen between the Pleasure-Dominance dimensions and the different aspects of the product's qualities. Non-significant negative correlation between the vocal assessment of Arousal and product's qualities (unlike the corresponding subjective self-reports) could be attributed to the inherent limitation of the self-report method, where participants may want to fulfil (or refute) the study's goal that they surmise.

Table 5.11 shows the correlations between the emotions derived from concurrent vocal expressions with the Automatic analysis approach and the product's perceived qualities. The only significant correlations found were between pleasure and goodness, and pleasure and classic aesthetics. Two correlations were also observed between pleasure and usability (r=0.34, $p < 0.05$) and attitude (r=0.35, $p < 0.05$) aspects.

AttrakDiff2	Pleasure	Arousal	Dominance
Pragmatic	0.14	0.22	0.33
Hedonic	0.20	0.14	0.29
Beauty	0.54**	0.14	0.41*
Goodness	0.73**	0.36	0.57**
Porat & Tractinsky	Pleasure	Arousal	Dominance
Usability	0.50**	-0.16	0.59**
Classic Aes	0.67**	-0.21	0.44**
Expressive Aes	0.59**	-0.22	0.48**
Attitude	0.63**	0.35	0.46**

Table 5.10: Spearman's rank-order correlations between derived emotions from concurrent vocal expressions with the Manual assessments and the product's (CompDealeres.com) perceived qualities based on the Attrakdiff2 and Porat and Tractinsky scales. **. Correlation is significant at the 0.01 level and *. Correlation is significant at the 0.05 level.

AttrakDiff2	Pleasure	Arousal	Dominance
Pragmatic	0.22	0.30	0.12
Hedonic	0.28	0.09	0.20
Beauty	0.23	0.02	0.04
Goodness	0.38*	0.01	0.06
Porat and Tractinsky	Pleasure	Arousal	Dominance
Usability	0.34	0.10	0.09
Classic Aes	0.37*	0.23	0.06
Expressive Aes	0.14	0.11	0.28
Attitude	0.35	0.14	0.10

Table 5.11: Spearman's rank-order correlations between derived emotions from concurrent vocal expressions with the Automatic assessments and the product's (CompDealeres.com) perceived qualities based on the Attrakdiff2 and Porat & Tractinsky scales. *. Correlation is significant at the 0.05 level.

These results to some extent confirm the correlations between the measures of emotions by Automatic and Manual assessments: While correlations were found between the two assessment methods in measuring pleasure and arousal, relations were also seen between product's qualities and the corresponding $\text{P}_{\text{Voc-Man}}$ and $\text{P}_{\text{Voc-Ac}}$. On the other hand, no relation was observed between product's qualities and the both corresponding $\text{A}_{\text{Voc-Man}}$ and $\text{A}_{\text{Voc-Ac}}$.

5.3.3 Effects of Context on UX

The RQ5 presented in this study sought to examine how the situation of an interaction influences the appraisal of the emotional experience and overall evaluation of an interactive product. To induce two different contextual conditions, participants were asked to read one of the fun- or task-related scenarios. According to the process of the study (Figure 5.8), the emotional status of participants were measured before and after reading the assigned scenario (by using SAM, each dimension in a 9-point scale). The idea was to verify whether the scenarios have induced changes in the emotional status of participants with respect to the given situation. In other words, reading scenarios was taken as a means of manipulation to induce task- or fun-related focus in participants.

In the context of online shopping, the driving force for an experience could be hedonic or utilitarian. The desires for gaining adventure (such as shopping for pleasure and stimulation) and gratification (shopping reliving negative feelings) are two important motivations in a hedonic shopping experience (Arnold and Reynolds, 2003). Hence, it was expected that the fun-related scenario could evoke the sense of pleasure and excitement in participants. On the other hand, achievement and efficiency have been identified as the driving keys in a utilitarian shopping experience (Babin et al., 1994). The task-oriented scenario was designed to explicitly emphasise accomplishing a task with an instrumental goal, highlighting the achievement purpose of the experience, whereas the fun-oriented scenario was designed to activate the sense of freedom and spontaneity.

The induction of the fun-related focus was accompanied by significant changes revealed by Wilcoxon Signed-Rank Tests. Significant differences were observed in the level of Pleasure (Z = -3.02, $p < 0.01$) and Arousal (Z = -3.24, $p < 0.001$), as measured by the pre- and post-scenario-reading SAM scales. For this group, the level of Pleasure increased from Mean = 6 (SD = 0.29) to Mean = 7.13 (SD = 0.30) and the level of Arousal increased from Mean = 4.75 (SD = 0.39) to Mean = 6.31(SD = 0.36). Although the level of Dominance slightly changed from Mean = 6.19 (SD = 0.38) to Mean = 6.56 (SD = 0.44), it was not significant and remained positive.

On the other hand, the induction of the task-related focus did not lead to any significant differences in emotional changes. However, there was a drop in the level of Pleasure from Mean = 6.43 (SD = 0.39) to Mean = 5.86 (SD = 0.39). The level of Arousal remained positive, changing from the Mean = 5.93 (SD = 0.39) to Mean = 6.29 (SD = 0.37). For Dominance, no significant change was observed either, where the Mean = 6.29 (SD = 0.43) slightly dropped to Mean = 5.86 (SD = 0.42). No changes in the level of Dominance after the induction phase could corroborate previous research, where changes in Dominance was not assessed in similar study's design (Hassenzahl et al., 2008).

In this regard, the context or situation of experience was the independent variable, and emotional experiences and the overall evaluation of a product's qualities (such as hedonic, aesthetics and usability) were the dependent variables. In other words, the research question was to explore how the context of an experience would affect the overall emotional appraisal when a product is being used.

		Fun-oriented	Task-oriented	
	Overall	Mean (Standard deviation)		MWU-test
AttrakDiff				
Pragmatic	5.19 (1.33)	5.09 (1.40)	5.30 (1.28)	n.s
Hedonic	4.72 (1.14)	4.79 (1.25)	4.66 (1.03)	n.s
Beauty	4.81 (1.14)	4.75 (1.25)	4.93 (1.03)	n.s
Goodness	4.72 (1.14)	4.80 (1.25)	5.00 (1.03)	n.s
Porat and Tractinsky				
Usability	5.62 (1.23)	5.40 (1.51)	5.86 (0.82)	n.s
Classic Aes	5.65 (0.64)	5.55 (0.96)	5.75 (0.71)	n.s
Expressive Aes	4.33 (1.25)	4.33 (1.37)	4.32 (1.71)	n.s
Attitude	4.94 (1.75)	4.61 (1.82)	5.30 (1.66)	n.s

Table 5.12: Mean (standard deviation) of the fun- and task-related groups on measuring Pragmatic and Hedonic qualities based on AttrakDiff2 (minimum-1 to maximum-7), and Usability, Classic and Expressive Aesthetics (Aes) and overall Attitude based on the Porat and Tractinsky (minimum-1 to maximum-9) scale. MWU-test (Mann-Whitney U test) column shows the significant differences between the two groups on their perceived qualities.

Effects of Contextual UX on perceived Products' Qualities

The mean and standard deviation of each of the measured constructs for the fun- and task-related groups are given in Table 5.12. While pragmatic (5.19/7), hedonic (4.72/7), usability (5.62/9) and classic (5.65 /9) aesthetics constructs were rated above the mid-point (positively), expressive aesthetics (4.33/9) was assessed below the mid-point (negatively) across all participants.

With regard to perceived qualities and their differences between the two groups, no significant variations were observed when computing Mann-Whitney U tests. Both groups rated the different qualities of the CompDealers website in a same range of slightly above the mid-point of each scale. The overall attitude towards the system was positive for the task-group, but negative for the fun-group, although this difference was not significant. In general, results suggested that having instrumental goal or not did not impact the perceived qualities and all participants evaluated the artefact relatively the same.

Effects of Contextual UX on Emotions

The influence of situation (fun or task) on the retrospective appraisal of emotions was investigated. As it was mentioned earlier, the retrospective appraisal of emotions was carried out by using the Porat and Tractinsky and DES scales (Section 5.3). Table 5.13 shows the mean and standard deviation of each measure of emotional appraisal for the fun- and task-related groups. On the whole, the experiences were perceived as emotionally positive and rather surprising. Significant differences between fun-

		Fun-oriented	Task-oriented	
	Overall	Mean (Standard deviation)		MWU-test
PAD				
Pleasure	5.97 (1.70)	5.61 (1.99)	6.37 (1.70)	n.s
Arousal	5.36 (1.25)	5.26 (1.80)	5.49 (1.62)	n.s
Dominance	5.87 (1.26)	5.77 (2.05)	5.98 (1.64)	n.s
DES				
Interest	3.76 (0.59)	3.65 (0.64)	3.79 (0.52)	n.s
Enjoyment	3.12 (1.16)	3.10 (1.21)	3.08 (1.15)	n.s
Surprise	2.71 (1.05)	3.04 (0.64)	2.36 (1.06)	n.s
Sadness	1.66 (0.76)	2.02 (0.64)	1.26 (0.38)	U = 39.5, p =0.002*
Anger	1.44 (0.82)	1.75 (0.64)	1.10 (0.24)	U = 68.0, p =0.029*

Table 5.13: Mean (standard deviation) of the two fun- and task-related groups on emotional appraisal based on the Porat and Tractinsky, minimum (1) to maximum (9) and DES, minimum (1) to maximum (5), scales. MWU-test (Mann-Whitney U test) column shows the significant differences between the two groups on their emotional appraisal.

and task-related groups emerged for emotional experiences of sadness and anger (negative experiences). While participants in both conditions were exposed to the same features of a product, the fun-related group evaluated their experiences more negative than their counterparts in the task group. This could be explained by this assumption that the fun-group had more expectation of gaining gratification from their experience than the task-group. However, a moderate experience of enjoyment (mean=3.10/5), pleasure (5.61/9) and surprise (3.04/5) by the fun-group, resulted in a more overall negative experience compared to the task-group. These differences can be taken as evidence for the successful manipulation of the induction phase. Accordingly, the fun-related group was set to attain pleasurable experience, but as the goal could not be fulfilled satisfactorily, negative emotional appraisal was resulted.

5.4 Discussion

5.4.1 Emotion Assessment Methods

Different assessment methods were used for measuring emotional UX, which could fall into the two components of model of emotions by Scherer (2005), subjective (self-reports) and expressive (vocal and facial). Using the subjective self-reports, concurrently and retrospectively, showed strong correlations in measuring emotions collected during and after the experience. In other words, the effect of memory biases was not observed in the results of this study. This could be explained by the fact that a number of peaks of emotional reactions had been evoked during the experiences.

Although it was intended by the inclusion of a short clip on the Homepage of the CompDealers website, the highest point of emotional responses was elicited at this phase. However, due to the failure in evoking one single peak, the most intense points were interwoven among the other emotional states, hereby resulting in correlations between retrospective and concurrent measures. The current finding is in line with previous work (Bruun and Ahm, 2015), where comparable variance in fluctuations had been reported in participants' experiences.

Regarding measuring emotions via the expressive component, strong correlations were confirmed between human ratings and automatic vocal analysis. In the related research by Zaman and colleagues (2006), the same trend of results was observed. In their study, using facial analysis by the FaceReader software as a concurrent measurement tool, measures obtained from FaceReader was very similar to the human ratings. Given the human capability of taking the context and the verbal content into account, associations were expected between subjective self-reports and human ratings. However, the only strong correlation was found between the measures of the subjective self-reports and human ratings was for the arousal dimension. On the other hand, according to Scherer (2003), there is an explicit distinction between expression and perception of emotions. In the vocal communication of the emotions, a speaker encodes his/her emotional state in the speech, whereas, a listener perceives the information transmitted by speech and makes inferences about it. As a result, the intended emotion expressed by the speaker may not necessarily match with the perceived emotion (Busso and Narayanan, 2008b). In other words, no strong correlations in measuring Pleasure and Dominance through the subjective self-reports and human ratings could be attributed to the complex problem of distinction between expression and perception.

The same underlying argument could be used to explain the fact that no associations between the automatic analysis and subjective self-report ratings was found. However, it is not surprising that the process of emotions recognition by automatic analysis is more complex than human perception. On the other hand, the overall trend of similarity between human ratings and automatic analysis highlights the practicality of vocal analysis as a concurrent measurement method, albeit requiring further improvements.

5.4.2 Vocal Emotion Analysis

Results of the data analysis showed the overall accuracy of 66%, 70% and 41% for each of the Pleasure, Arousal and Dominance dimensions respectively. Recognising the emotional content of natural data is a challenging task, especially considering background noise being captured during the recording time (Marchi et al., 2015b).

However the recognition performance for Pleasure and Arousal is considerably and for Dominance is moderately higher than the chance level. This outcome could suggest the effectiveness of the methodology used in this study.

5.4.3 Facial Emotion Analysis

Facial signals are reliable source of data for recognising emotions. The application of facial expressions in usability test has been tested in previous work (such as Zaman and Shrimpton-Smith, 2006). In this work, the Kinect sensor was used to capture facial reactions during system interaction and thinking aloud. A training set was created based on the concurrent self-reports of dimensional emotions and Kinect's AUs to examine the feasibility of using the Kinect for automatic assessment of facial expressions of emotions. The size of this dataset in terms of the number of collected instances is relatively small (N =111), however it encourages further research to record more data from Kinect in order to build a dataset for emotional analysis of facial expressions. Similar approaches to vocal analysis can be applied for assessing emotions from facial reactions.

Results of regression analysis suggested that particular AUs outputted from the Kinect contributed to the prediction of the PAD dimensions. Whereas previous work (such as Mehu and Scherer, 2015) employed human coding for describing AUs, the present work made use of the inexpensive Kinect sensor, yet demonstrating same correlations between AUs and emotion dimensions (such as the negative association between arousal and brow lowering). This finding supports the feasibility of using the Kinect for evaluating emotional expressions through facial movements.

Despite the potential capability of Kinect for describing facial movements, the current version of this sensor can only capture a subset of AUs compared to the original FACS. Consequently, a comprehensive analysis of facial movements from Kinect is still not possible.

5.4.4 Contextual UX

According to previous research, having instrumental or non-instrumental goal has impacts on retrospective appraisal (Hassenzahl et al., 2008). However, perceived quality is a consequence of both the product's character and the situational aspect. In other words, whether a product is primary hedonic or pragmatic in various situations (fun or task) would shape different appraisal with respect to the perceived quality. In this study, the interactive product was neither primary pragmatic nor hedonic (Table 5.12), in other words a balance of both qualities was considered. Hence, the context of the use alone did not have an effect on the perceived quality.

On the other hand, significant differences were observed in emotional appraisal between the two groups. Experiencing pleasurable experience is an important goal for fun-seeking group. However, failure in fulfilment of this goal led to rather disappointment or negative emotions (Table 5.13) for the fun-seeking group, which in turn resulted in a poorer attitude of this group (4.61/7) towards the system compared to the task-oriented group (5.30/7). Overall, the present results supported the importance of the context of the use in appraisal of an experience, which is not stable, and depending on situation changes.

Another interesting outcome was the differences beheld in the findings of the two instruments: The dimensional PAD, managed by bipolar items and the discrete DES which was administered by unipolar items. While the both instruments are used for measuring emotions, only through the DES discrete emotion scale, the differences in emotional appraisal of the two fun- and task-group were inferred. This could be traced back to the possible effects of bipolarity versus unipolarity of an evaluation. Some researchers such as Kaplan 1972 advocated that positive and negative evaluations can be independent and thus not distinguishable from the bipolar scales. In an ambivalent appraisal, while some aspects of an experience can cause a negative perspective, other aspects can result in a positive evaluation. Therefore, using the bipolar PAD scale did not reveal the differences in emotional appraisal between the the two groups. In addition, Table 5.1 showed a negative and strong correlation between arousal and anger. It means when the level of arousal was increased, anger was decreased. This result is in contrast with the literature, where anger is usually understood as a high arousal state. In other words, the items of the PAD scale failed to indicate anger with high arousal.

Moreover, the significant correlations existing within the PAD dimensions of the self-reports of emotions (Table 5.1), for instance between pleasure and arousal, indicate participants' tendency to report the positive pleasure dimension with high arousal and conversely negative valence with low arousal. On the contrary, in the work of Bradly and Lang (1994), low and non-significant correlations were observed within the PAD dimensions when using SAM and the Semantic Differential scale. However, depending on the culture in which a study is conducted, positive or negative relation between pleasure and arousal was observed by previous work (Russell et al., 1989). Consequently, given the evidence of the possible relations between those dimensions, it is justifiable to assume that participants in this study also associated high arousal with positive pleasure.

In other words, the emerged strong and positive correlations within the PAD dimensions could imply that evaluation of particular emotional experiences such as anger may not be supported by the PAD scales, when they are used to assess emotions in the related studies. As a result, significant differences between the fun- and task-oriented groups were not attainable using these instruments. In contrast, employing a categorical emotion measurement scale such as DES revealed such differences between the two groups.

Nonetheless, the given finding does not suggest the superiority of the instruments consisting of discrete emotions over the PAD-based instruments, but it mainly highlights one of the advantages of instruments based on the categorical labels over those with dimensional descriptors.

5.4.5 Threats to Validity

The results of the presented study are subject to threats of validity. Repeat testing, as it was observed in evaluating H2.1, affected the outcome. Having previously taken part in Study 1 led some of the participants to make assumptions of the purposes of the current experiment, and hence to influence their more natural responses. Another major validity pertained to the instruments were used, where the differences in emotional appraisal of the two fun- and task-group were only indicated using the DES scale and not the PAD. In other words, these two instruments did not produce the same outcome from participants' data.

In terms of the methods used for assessing the associations between the concurrent and retrospective experiences, positive and strong correlations were found in measuring Pleasure and Dominance where subjective self-reports and manual ratings were employed. On the contrary, no association was verified in measuring Arousal using automatic and manual ratings. While the discrepancies in the results could advocate for the advantage of a method on another (e.g. assessing Arousal using expressive signals instead of self-reports), yet these inconsistencies could be related to the degree of the accuracy of the method in its measurement.

Similar to Study 1, the generalisability of the results can be questionable due to the sample of participants and experimental setup. In other words, different settings and group of subjects could lead to different conclusions.

Chapter 6

Conclusion

This chapter first presents the general discussion on the main findings and implications of this thesis for research and practice on three areas: Think aloud, UX and vocal emotional analysis. Next, the research questions will be revisited. Furthermore, limitations and directions for future work will be covered.

6.1 General Discussion

6.1.1 The Application of Think Aloud Protocol

The present work proposed a methodological approach to measure UX based on the think aloud technique and by taking advantage of emotional analysis of vocal and facial expressions elicited during thinking aloud. The proposed methodology is grounded in one of the most used and cost-effective usability evaluation methods the think aloud protocol. The common practice of the think aloud approach is that an experimenter observes participants' behaviours and listens (records) in on their thoughts while they think aloud (Hertzum and Holmegaard, 2013). While the observation part provides information about the behavioural as well as non-verbal reactions and could be facilitated by the use of other techniques such as eye tracking, video recording and note taking, the verbalised content adds information about why, in other words the reasons and how participants perform a task and the explanation of their actions (Section 2.2.1).

Coding of Think Aloud Utterances

Employing techniques such as eye tracking simultaneously with thinking aloud has further helped earlier research with understanding users' cognitive processes during usability testings (Cooke and Cuddihy, 2005). In other words, observational techniques has called attention to the limitations of think aloud protocol, where subtle relevant user data such as locating specific information on an interface or emotional responses could be easily missed (Cooke and Cuddihy, 2005). Aligned with the same argument, it could be claimed that emotional responses conveyed via vocal and facial expressions are not collected by the current techniques in thinking aloud sessions. On the contrary, analysing this source of data manifested from expressive reactions could enhance the value of the user-based studies evaluations. For example, the source of information could help designers to understand when and why users are emotionally or cognitively more involved.

Unit of Analysis

Segmentation is a common preprocessing step, which must be carried out for both vocal emotional analysis (as it was described in Sections 4.2.4 and 5.2.5) and thinking aloud's content analysis. This task can be performed based on objective criteria or semantically meaningful units. Hence, finding a standard procedure would speed up the process of data preparation in both of the methodological approaches. While relying only on objective criteria could result in masking meaningful units due to the fact that automatic segmentation has the possibility to span over more than one semantic unit (Schuller et al., 2011a), pure manual segmentation has the drawback of introducing huge cost of human labour. Clearly, there is a trade-off between precision and human effort invested. This cost-effectiveness issue is largely dependent on the precision level that a particular application requires (e.g. analysing the emotions of a suspect's interview may demand a higher precision than a usability test participant's thinking aloud.)

The think aloud protocol was originally developed for getting access to the cognitive processes. However as it was described, the recent use of this protocol shows the application of this method for emotional assessment as well. Given the more reliability of segmentation based on semantic units (Chi, 1997), the questions could be raised whether the emotional and cognitive content remain consistent simultaneously throughout an utterance segment? Or whether segments convey mainly either an emotional (such as related to negative and positive experiences) or a cognitive (such as an action and the related reasoning) meaning? According to the appraisal theory (Scherer et al., 2001), people's emotions are elicited as result of their evaluations or appraisal of events and situations. This means that emotions consist of perception and interpretation of circumstances, which involve cognition. In addition, manifestation of emotions through the expressive component is known to be the adaptive externalisation of cognitive appraisals and their linked action tendencies. Although there are times that thoughts are entirely free of emotions (for example being in neutral state), thinking and feeling are interrelated most of the times (Ellsworth and Scherer, 2003). Therefore, segmenting verbalisations based on the criterion of "emotionally consistent meaningful units" could imply that each segment is a thought without (neutral) or with one emotion throughout.

To balance the advantage and disadvantages of the automatic (such as silence pause) and manual segmentation approaches, a combination of both was adopted in this work. In this regard, thinking aloud utterances were first segmented based on the silence pause as an objective criterion, then two human coders evaluated them according to the notion of "semantically meaningful units with no change of emotion" as the criterion. As a result, the ultimate boundary for utterances was defined, where both evaluators were in agreement.
The average numbers of words per utterance were 10.31 and 11.4 respectively for the Pilot and Main studies in this research (Sections 4.2.4 and 5.2.5). In addition, utterances in IEMOCAP corpus composed of 11.4 words on average. As these numbers show the average number of words per utterance has fallen in the same range (10 to 12). Given these empirical results, for future work an automatic segmentation could be performed in terms of the number of words in an utterance. However, more research is required to substantiate the current findings.

Number of Participants

This work involved two think aloud studies with 46 and 35 participants. While these sample sizes seem higher than average (Table 2.1), a higher number of participants could ideally be recruited for the analysis based on the ML techniques. The recruitment, however, proved challenging, especially given the time constraint of the studies. Nevertheless, the resource-demanding qualitative analysis (such as coding) and preprocessing steps (such as segmentation) could also justify that the number of participants in think aloud tests compared to survey-based (e.g. questionnaire) studies, where the number of respondents are usually higher.

Finally, concurrent think aloud was carried out in this research due to the findings of previous work, where more problems were identified by the means of observational data in concurrent think aloud as opposed to the retrospective approach (Van Den Haak et al., 2003). In addition, concurrent think aloud has the advantage of incorporating the *in situ* aspect of evaluation, which enables collection of contextual momentary experiences rather than a retrospective report of them. On the other hand, research has shown trade-offs between the two approaches in terms of the number of utterances as well as the extent of explanation in verbal contents (Van Den Haak et al., 2003). Hence, the comparison between the two think aloud protocols concerning automatic analysis of emotional assessments is essential to draw solid conclusions.

6.1.2 Implications for UX Practices

This thesis investigated emotional experiences with respect to:

1. An automated approach based on the techniques adopted in emotion recognition (Sections 3.3 and 3.4.3).
2. A moment-based approach as opposed to the memory-based assessments (Section 4.1).

Each aspect will be discussed in the subsequent sections.

Automated Approach for Measuring Emotional Experiences

The present work proposed and tested a methodology grounding on recognition of emotions based on both vocal and facial expressions to assess emotional experiences. Based on this methodology, a concrete application can be created to provide automatic feedback, detect and mark the cognitive and emotional incidents (such as when a user is being irritated by a usability problem) during the interaction. In this regard, the application can be served as a toolkit to automatically generate a summary (such as visual depiction) of the entire interactive experience and its key aspects, and thereby to eliminate the tedious work of manual analysis.
Assessing users' emotional and behavioural responses during usability testing has already been adopted using different techniques such as the survey-based approaches, log-files analysis of mouse and keyboard actions and eye-tracking technology. Consequently, it is also plausible to envision that advancement in automatic recognition of vocal and facial expressions can facilitate data analysis across the similar domains. Real-life applications of automatic analysis of emotions have been extensively explored for interactive environments, products and services. Examples of those application areas include ambient intelligent environments such as classrooms (D'mello and Graesser, 2007), meeting summary generation applications and augmented call centre services (Bedoya-Jaramillo et al., 2013). These examples acknowledge the current trend within the design of interactive systems for automatic integration of users' emotional states, thereby improving UX (including the pragmatic and hedonic aspects) of the interaction. In this regard, an automatic assessment tool could also be viewed as another example within the current trend of technology incorporating recognition of emotions.

Moreover, usability tests have been conventionally conducted by a moderator within a laboratory's setting. On the other hand, unmoderated usability tests have been increasingly becoming more popular (Liu et al., 2012). In unmoderated usability tests, the users run the test session themselves, most likely in their homes, while their utterances and behaviours are recorded for subsequent analysis. Although crowdsourcing has reduced the cost required for conducting usability testing, assessing verbalised and non-verbalised data is yet another resource-demanding task in terms of time, labour and cost. Consequently, automated techniques for analysis alongside crowdsourcing can complement the reduction in the cost of evaluation. Especially, automatic analysis could be of particular concern to UX studies, where current evaluation methods tend to collect data longitudinally (e.g. Kujala et al., 2011). As a result, a huge volume of data will be available to be analysed which could be eased by automatic processes.

To conclude, an automated approach for assessing emotional experiences can replace the traditional techniques in terms of collecting and analysing data, therefore to facilitate UX practices.

Momentary vs. Retrospective Assessments

A compelling design is driven by understanding people's experiences. This understanding can be framed based on objective or subjective measures collected ahead of, in the moment and after the experience. In particular, the temporal dependency of measures can lead to different impressions of experience, where moment-based and retrospective measures do not always portray the same view of experience due to some memory biases such as the peak-end effect. Results of the Main study (Chapter 5) showed significant associations in the measures of Pleasure, Arousal and Dominance between the concurrent and retrospective assessments using self-reports.

Using vocal analysis as a moment-based method instead of self-reports supported corresponding correlations in measuring Pleasure and Dominance and not Arousal (Table 5.6). A similar finding was observed in the results of the Pilot study (Chapter 4), where the only correlation was found in measuring Dominance. Therefore, in both studies, vocal and momentary measures of Arousal did not associate with the retrospective self-report of this dimension. In other words, a gap was observed in measuring Arousal between the retrospective self-report and the vocal analysis (as a moment-based measure) of this dimension.

The reported inconsistencies in relations between the retrospective (self-reports) and momentary (self-reports and vocal analysis) measures of Arousal could be explained along three directions. First, the interactive systems used in the experiments were not emotionally stimulating enough to cause changes in Arousal that could be perceptible through vocal expressions. Second, the adopted acoustic analysis failed to capture the emotional expressions through voice. Third, the retrospective self-report of Arousal did not indicate the actual experience of this dimension due to the memory-gap effect. The significantly lower self-reported measure of the retrospective Arousal (mean= 5.36) as compared with the concurrent Arousal (mean= 6.11) in the Main study (Section 5.3.1), lends evidence to the third assumption. However, the first or second assumption could have also played a role.

Moreover, the gap between retrospective and momentary measures was only seen in assessing Arousal. This signifies the advantage of using the dimensional schemes, where each emotion is modelled along three independent descriptors instead of one individual label. Results of this study suggested that the memory-gap effect may not affect all the PAD dimensions to the same extent or even at the same time. One implication is that depending on the context or a product/service's design features, momentary measurements are more relevant for evaluating users' emotional experiences..

6.1.3 Vocal Emotion Analysis

Emotion recognition from voice is a field that is receiving more and more attention from researchers in speech recognition as well as in the HCI community. The overall approach of voice emotion recognition hinges on the machine learning techniques to train and develop models based on acoustic features extracted from corpora of emotional speech. Although the general architectures of the systems are similar, there are variations in the implementation methodology of each of the components. Diversity exists in data collection methods, such as the emotion elicitation techniques and labelling approaches, calculation and selection of relevant acoustic features, and employing different machine learning algorithms for recognition of emotions.

Given the dissimilarities among the different components, recognition results are hard to be compared across different studies. However, call for standardisation has been made within different conferences such as the series of INTERSPEECH Challenge, where for instance, baseline feature sets and a training dataset are usually provided for participants to present their results. In order to recognise emotions elicited during thinking aloud in this work, the general procedure followed by the INTERSPEECH Challenges (from 2009 to 2017) has been employed. For example, extraction of the acoustic features relevant to emotions is carried out through the use of the openSMILE signal processing software and its integrated emotion related feature sets such as the INTERSPEECH 2009 and GEMAPS feature sets.

To clarify the implications of the different approaches taken for each of the components of the voice emotion recognition system in this work, the following sections will outline the possible variations in the applied methodology and the relevant issues to be tackled.

6.1.4 Corpora Selection

As explained before (Chapter 3), performance of emotion recognition from voice is highly dependent on the corpora being used for data training. Emotional speech corpora can be roughly divided into two camps acted (simulated) and natural (spontaneous) data. Despite the number of acted corpora, acted emotions cannot fully correspond to the emotions displayed by regular people in real-life situations. In real-life scenarios, a mixture of emotions with different intensities elicited as opposed to the exaggerated acted emotions (Douglas-Cowie et al., 2003). As suggested by the appraisal theory, emotions are exhibited as a reaction to events. Hence a contextual scenario is required to trigger emotions, whereas acted corpora are usually non-interactive speech and without a scenario to trigger emotions. Consequently, some researchers have argued that the elicitation method is the main attribute which makes acted data not fully representable for real-life situations and not the use of actors per se (Busso and Narayanan, 2008a). Given these rationales, this research opted for the use of IEMOCAP corpus as the main training dataset due to the reason that this corpus contains emotional scripts and improvisations of fictitious situations. In other words, the combination of the fictitious situations to elicit emotions and the interactive aspect of IEMOCAP made this database a suitable training set for emotion recognition of think aloud utterances. The fictitious context would correspond to eliciting emotions in real life situations and the interactive aspect would be related to to the expressed emotions in a communicative setting.

On the other hand, research has shown that with the introduction of multiple corpora, improvements on the performance can be expected. Especially since classification models that are generated on a set of data are usually database-dependent and cannot be generalised well on new and unseen type of data (as it was experimented as well from the results of the analysis in Chapter 3) (Lefter et al., 2010). Due to the disadvantages of using acted emotional data, efforts on recording natural emotional corpora have been increased despite the challenges on recording such data (such as the ethical and copyright issues). Examples of natural emotional corpora are given in Tables 3.1, 3.2, 3.3 of Chapter 3. Many of the natural emotional databases have limited (or with a licence fee accessibility) for public use, as it was discussed in Chapter (3). Language of recording, which is another barrier of using different corpora, has also a limited variation when it comes to natural emotional data.

6.1.5 Emotional Descriptors

Both categorical and dimensional labels provide useful complementary information. For example, dimensional attributes can be used to indicate the intensity of the segments labelled with the same emotional category (Busso et al., 2013). On the contrary, there are limitations in using either of the models to build an emotion recognition system. Within the discrete categorical model, clear distinctions between different emotion labels is not always obtainable, which can result in a lower inter-rater agreement on emotion categories as compared to the dimensional labelling (e.g. in Busso et al., 2008). On the other hand, there are a fewer number of corpora annotated based on the dimensional models (as it was seen in Tables 3.1, 3.2, 3.3). To overcome the aforementioned constraints, discrete emotional categories were translated into the dimensional descriptors, based on the theoretical background that various emotional states are definable as regions in the three dimensional structure (Russell and Mehrabian, 1974). Taking this approach, first of all, helped the process of manual annotation of the think aloud utterances (given the blurred distinctions between emotion labels). Second, mapping emotion labels into dimensions provided a common space to compare and combine corpora with different categories of emotions. For example, in a cross-corpus evaluation, previous work only considered the labels that were present in all the used datasets (Lefter et al., 2010).

However, it is noteworthy to mention that the three dimensions could contribute unequally in defining different emotion categories (for example the degree of displeasure has been suggested twice as important as arousal in defining anger (Russell and Mehrabian, 1974)). Therefore to convert emotion categories into dimensions, these variances should be taken into account.

According to research (Busso et al., 2013), the emotional labels should be driven by the application at hand. For UX evaluation, emotions are usually categorised and assessed based on the positive and negative labels, such as the the emotion labels in the PANAS scale. Another extensively used method for describing emotions for UX assessment is based on the dimensional emotions, such as the use of SAM scale, where emotions are characterised according to the three dimensions Pleasure, Arousal and Dominance. Consequently, a combination of these two theoretical concepts of describing emotions was adopted in this work. Accordingly the combination of positive versus negative and three dimensions led to an approach for the classification of the emotional content of verbalisations to the three dimensions Pleasure (Negative/Positive), Arousal (Low/High) and Dominance (Low/Neutral/High).

While neutral state was categorised as low Arousal and positive Pleasure, a separate category was determined for evaluating this state in terms of Dominance. This consideration was due to the lack of empirical research on recognition on this dimension. Many recognition systems have focused exclusively on two-dimensional models including Pleasure and Arousal (e.g. Marchi et al., 2015a; Schuller et al., 2011c). However, the current work presented results involving the third dimension, Dominance or Potency, as well. Therefore in order to avoid the negative effect of falsely adding neutral states to the either class of low or high Dominance in a binary classification, instead, classifiers were trained for prediction on the three-class problem of low, neutral and high.
Moreover, although following the previous research, neutral states were added to the low class of Arousal and positive Pleasure, theoretical and empirical research is essential for such clustering scheme in order to build accurate recognition systems. The binary clustering approach for Arousal and Valence was particularly adopted in order to improve the recognition accuracy through: 1) An easier classification task by discriminating between two classes instead of multiple classes. 2) Handling the problem of imbalanced distributions amongst classes.
It should be noted that a binary decision classification task could be extended to subsequent levels of multi-class decisions in a hierarchical structure (Lee et al., 2011). This approach will be investigated in the future work.

The Dominance dimension is of particular importance for distinguishing emotional states such as anger from anxiety (or fear), where the former is associated with high dominance and the latter is a submissive state (Russell and Mehrabian, 1974). Moreover, inclusion of a fourth dimension, Predictability, has been proposed to allow description of reactions towards a novel stimulus or an unfamiliar situation (Fontaine et al., 2007). According to Fontaine and colleagues the fourth dimension is particularly associated to the state of *surprise* in emotional experiences. Surprise is one of the six basic emotions (Ekman, 1992) in the discrete model of emotions and can be used for the appraisal of novelty or unexpectedness of events or situations.

It should be noted that analysing emotional reactions in terms of predictability of events is especially intriguing for UX research, where novelty and stimulation are perceived as product/systems' hedonic attributes (Hassenzahl, 2003). Consequently, the future work of this research will be extended to measure the Predictability in addition to the PAD dimensions. However, similar to the PAD dimensions and categorical labels that extensive research has been carried out in terms of changes expressed in vocal and facial features (such as Mehu and Scherer, 2015; Scherer,

1986), related research is also required for the analysis of the Predictability dimension.
To build a system for recognition based on the four dimensional model, one approach is to use the current emotional corpora and map their discrete labels to the fourth dimension in addition to the PAD dimensions. However this task could be facilitated by development of corpora adopting the four dimensional model. An example of such corpora is the MAHNOB-HCI multimodal database (Soleymani et al., 2012). MAHNOB-HCI contains video clips of facial reactions accompanied by eye-gaze recordings and physiological responses, which has been annotated against the four emotional dimensions.

Given the PA or PAD dimensions cannot fully represent the differences and similarities between the emotional states, it could be implied that the basic model of emotions can better reflect the dissimilarities between the expression of emotions. This argument could be supported by the results of the Main Study (Table 5.13), where the differences between he two groups of the presence or absence of task-related goals were only identified by using the DES scale. In other words, the use of folk language of emotions could better help understanding emotional experiences.
However, in conclusion I would like to argue that both models of emotions can derive valuable insights about emotional experiences. For example as it was discussed, the effect of memory-gap was only observed in the measures of Arousal, where there was a significant difference between the concurrent and retrospective measures of this dimension. This result indicates the fact that each dimension carries specific information about emotional states. A specific example could be given about Dominance, where action tendency differs between the state of being apathetic versus controlling (Fontaine et al., 2007).

Feature Selection and Classification

One of the crucial steps in vocal analysis of emotions is the extraction of the most emotion-related features set that is reasonably concise (Schuller et al., 2011a). As it has been explained throughout this thesis, prosodic, voice quality and spectral features of voice are among the most important features for recognising emotions. Since these features are complementary to each other, a combination of them has resulted in improving the performance of the recognition systems compared to those systems developed using only individual features (examples of the research studies can be found in the review by Koolagudi and colleagues (Koolagudi and Rao, 2012)). To identify what features and to what extent they contribute to recognising emotions, extensive auditory research has been carried out since decades ago (such as Scherer, 1986; Schuller et al., 2007a). While early studies used hand-crafted features using tools such as Praat, since the INTERSPEECH 2009 Emotion Challenge, expert-knowledge based feature sets have been utilised for automatic extraction using toolkits such as openSMILE (e.g. Ringeval et al., 2016)). The motivation of this choice is: First, openSMILE with integrated feature sets is freely available and this combination provides a well-defined standard for emotion recognition especially on real-life related applications (Marchi et al., 2016). Second, the fact that by this combination, one can reproduce the same benchmark results in terms of recognition performance.

Consequently, due to the advantages of employing openSMILE, the present work was also grounded on this technique for the extraction of the acoustic properties relevant to vocal emotion expressions. Later, the extracted features were fed to SVM classifiers for the prediction task on emotions. On the other hand, the baseline acoustic features can be used as the input feature set to Recurrent Neural Networks (RNN) classifiers (Hochreiter and Schmidhuber, 1997). RNN classifiers are capable of making prediction decisions considering contextual information. In this regard, a decision at time step *t-1* influences the decision at time step *t*. In future work, the RNN techniques will be investigated for taking the temporal information of trajectory emotions and as methods for comparison with the current employed approaches.

Two classification methodologies were used in this work: One methodology was based on the conventional classification, where the prediction model is trained on a dataset and it is tested on unseen data (as it was illustrated in Figure 4.4). This approach has been used in various studies for classification of emotions until recently (such as Ringeval et al., 2016), despite a number of limitations: First, models generated through the conventional approach usually result in poor generalisation when performing on new data. To address this issue, one approach is to leverage classification models by limited amount of data from the unseen or target domain, thereby enabling models to be generalised better to new settings. Convolutional Neural Networks (CNN) have shown good performances for model adaptation. Second limitation concerns considering contextual information as explained in the above paragraph. Accordingly, emotion recognition on think aloud utterances could be optimised alongside tackling two aspects: adaptation and contextualisation. Future work will investigate Deep Machine Learning (DL) techniques such as RNN and CNN.

6.2 Insights into the Research Questions (RQs)

RQ1: *Whether changes as expressed through think aloud verbalisations are sufficient enough to be measured/detected to indicate the corresponding changes in the quality of interaction with a service/product?*

The results of this research suggested fluctuations in emotional experiences during a course of interaction. These changes were observed between the verbalised contents by human raters and the analysis of acoustic features. In other words, the changes in emotions were large enough to be manifested in voice and therefore be detected. In addition, significant correlations were confirmed between the measures of momentary emotions assessed by vocal analysis and qualities such as the perceived beauty (aesthetics) and usability of the system, thus the overall attitude towards it. Hence, the answer to RQ1 is affirmative.

Due to the close relationships between RQ2 and RQ3 and to automatic techniques for emotion detection, the answers to these research questions are presented as one section:

RQ2: *How can automatic techniques support the recognition of emotional expressions during thinking aloud?*
RQ3: *What are the requirements of building an automatic system dedicated to recognition of emotions from thinking aloud utterances?*

Given the emotional responses expressed through thinking aloud verbalisations, automatic approaches such as vocal and facial expressions analysis, linguistic and eye-tracking techniques can be used for recognition of emotional experiences. As it was discussed in Section 6.1.1, observational techniques (such as video recording) have already been in use for acquiring additional information about users' behaviours and cognition while thinking aloud and performing a task. Similarly, extension of the data analysis supplemented by verbal and nonverbal emotional responses (such as face and voice) can give insight into users' momentary experiences.
Advancements in automatic techniques imply better performances in terms of accuracy in recognition results. In other words, the performance of emotion recognition during thinking aloud is directly linked to techniques applied for any other recognition systems such as detection of emotions in tutoring or call centre applications. Therefore, proliferation in labelled data, precise extraction of features from a signal and advancement in ML techniques will benefit automatic recognition of emotional expressions during thinking loud.

The building blocks of an automatic system dedicated to recognition of emotions from thinking aloud verbalisations is similar to other recognition systems, however with some specifications. As it was mentioned in Section 6.1.3, the training corpus for model development has a significant role in the performance of a recognition system. Therefore, dedicated corpora to usability tests, with spontaneous data annotated according to the type and intensity of those emotional responses elicited during system interaction would be required. Next is the use of appropriate ML techniques that are: 1) Robust and generalisable to future and unseen data in order to tackle variations in recording conditions and individual differences in terms of expressing emotions. 2) Adaptable and capable of taking into account contextual information as emotional reactions are triggered as responses to the encountered situation and that can change rapidly.

Results of the data analysis of the Main Study (Chapter 5) showed significant correlations between measures of emotions where they were assessed manually and automatically by means of vocal cues (Figure 5.12). This indicates the feasibility and reliability of automatic analysis for recognising vocal expressions of emotions elicited during thinking aloud. Concerning facial analysis, employing an automatic technique by the use of Kinect suggested strong correlations between facial expressions and the subjective self-reports of emotions. Consequently, the answer to RQ2 is that automatic techniques can be employed for the recognition of emotional expressions during thinking aloud, although more research is required to improve the reliability of such techniques.

RQ4: *How can an automatic system facilitate the demanding process of analysing UX data?*

Qualitative methodologies are the main approaches for obtaining insight into UX (Bargas-Avila and Hornbæk, 2011), despite the high costs required in terms of time and labour to analyse such data. The think aloud protocol is one of the approaches that can be used for gathering qualitative data. Verbal protocols and in general given reviews are usually analysed manually and based on a coding scheme designed according to the objectives of the study. On the other hand, both automatic acoustic and linguistic analysis can substitute manual coding in order to categorise reviews based on their emotional and cognitive meaning as well as to weight their severity (for example by assigning a value as a level of interest to an utterance). Having carried out the process of analysis by automatic approaches will reduce the cost of manual labelling.

RQ5: *How are users' moment-by-moment emotions elicited when interacting with a* *service/product integrated into their aggregated emotion and thereby attitude towards the service/product?*

The present work aimed to propose a methodology that enables us to view UX with blocks of momentary emotions as well as from a retrospective perspective. In the procedure of this methodology, the momentary emotional experiences were detected and classified into different emotion classes in terms of categorised labels (such as anger or happiness) and dimensions (negative or positive). Results of the Pilot Study (Chapter 4) suggested that the emotion label with the highest number of occurrences or the *Modal emotion* could be used as a predictor for an attitude towards a product/service. Findings of the Main Study (Chapter 5) showed that aggregated momentary emotions described by the two dimensions pleasure and dominance are significantly associated with the goodness and attitude towards the system (Table 5.10 and Table 5.11).

6.3 Limitations

There are several limitations in this work, which have mostly been caused by the restricted resources available and the organisational constraints. Each of them will be described in the following:

Sample Size and Scope

The sample in this work consisted mainly of university students aged between 20 to 35 years old. Therefore, the generalisability of the results is restricted to university-educated people within this age group. Additionally, psychophysiological studies typically demand a large sample-size to derive reliable conclusions, especially due to the large individual differences in emotional reactions. Consequently, studies comprising larger and more divers group of participants would be useful to further substantiate the findings presented in this thesis.

Variations of Think Aloud

Concurrent think aloud was carried out as the elicitation method for emotional expressions during experiences. To examine the effectiveness of the think aloud technique for understanding emotional experiences, especially with regard to a holistic view, concurrent verbal protocols should be compared and evaluated against retrospective verbalisations. However, due to the large amount of data collected during thinking aloud and the subsequent pre- and post-processing data analysis (such as segmentation, manual coding, etc.), this work was limited to the concurrent version of this protocol.

In addition to the comparison between the concurrent and retrospective protocols, moderated and unmoderated thinking aloud should be investigated in order to fully address the main research questions in this study. This comparison is particularly essential due to the fact that the quantity and quality of expression of emotions can differ depending on whether an audience is present or not (Pennebaker et al., 2015). Therefore, it would be intriguing to find out how the results would vary without the attendance of the experimenter in a test session.

Automatic Analysis

As it was explained in the discussion section, the automatic analysis of vocal emotions was carried out based on the predefined feature sets describing emotions, using openSMILE for feature extraction and WEKA's implementation of different ML algorithms. Although this approach provided a straightforward and transparent procedure, which can be replicated for benchmarking analysis, it is limited to the extent of the employed ML techniques. Hence, future work is required to substantiate the analysis by investigating a broader scope of ML techniques.

Moreover, the current extent of analysis did not investigate recognition of emotions from the other spoken languages. In other words, the design of the recognition system in terms of training (corpora selection) and testing (participants) was restricted to English as the spoken language. The rationale behind this choice of design was related to the findings of previous research, where different recognition rates were observed when a classification model was used for different languages (Rajoo and Aun, 2016). In addition, this research was conducted in the United Kingdom. Rajoo and colleagues (Rajoo and Aun, 2016) showed that recognising emotions has a higher accuracy of prediction on emotions expressed by native speakers.

A comprehensive UX measurement method should be able to afford cultural differences, where people's expressions and perceptions of their experiences are likely affected by their culture (Ekman and Scherer, 1984; Law and van Schaik, 2010). Hence, the future focus of this research is to explore approaches for voice emotion recognition, which are independent of the underlying language. In this regard, first it is essential to determine whether there are specific features that exclusively contribute to a particular language. Second, ML techniques (such as Transfer Learning algorithms) should be investigated in order to test the efficiency of the classification algorithms, when they are trained with a set of limited labelled data. Consequently, a recognition system trained with instances of specific language such as English could be used for recognition in other languages.

With regard to facial expressions, the research was limited to measures of facial movements computed by Kinect. In other words, manual assessment of facial expressions was not carried out. This was due to the reason that manual coding of facial expressions requires specific training and it is very time consuming; coding of one hour of video data could demand four hours of work (Burr, 2006).

It may be argued that similar to the vocal analysis, training and classification techniques could have been applied for recognition of facial expressions tracked by the Kinect sensor. In this research, Kinect was used for the extraction of facial expressions based on a subset of AUs (similar to the use of openSMILE for extracting vocal features). Consequently to train classification models for recognition of facial expressions, feature extraction of training labelled data has to be performed as the prerequisite step.

One approach to bypass the process of feature extraction is to exploit emotional databases of the set of AUs corresponding to emotions in addition to the raw data (such as static images or videos for face and audio files for voice). An example of such dataset would be a larger extent of the data acquired in the Main Study (Section 5.3.1) of this research.

Recording of databases of facial expressions using Kinect has been reported (e.g. Aly et al., 2015), although with variations in the type of data being captured (such as the face location in a 2-dimensional image instead of the AUs). Therefore further research on the use of such databases for classification of facial expressions of emotions should be considered.

Experiential Domain

The current research investigated the concept of emotional experiences only in one domain of interactive application, online shopping. Thus, this undermines the generalisability of the findings to the other domains such as gaming, which is likely to elicit a (much) broader range of emotions. However, it is plausible to envision the applicability of emotional analysis of expressive emotions (vocal and facial) in natural conversation settings such as learning, entertainment and consumer dialogue.

Verbal Analysis

The contents of the verbal protocols were not analysed and compared against nonverbal data. While manual coding can be adopted for the content analysis of thinking aloud utterances (as described in Section 6.1.1), ML techniques can be used for sentiment analysis (opinion mining) of the verbal contents. In this regard, sentiment classification using supervised learning algorithms (such as SVM) can be applied to classify utterances based on positive and negative classes. However similar to any ML approaches, extracting a set of relevant features is required prior to classification. One direction of the future research will be on the application of automatic techniques for the content analysis of verbal data. Subsequently the results of automatic analysis can be compared with the manual coding of the contents.

6.4 Conclusion and Future Works

Figure 6.1 proposes the design of the required components of a UX assessment tool based on the think aloud protocol. Given the implications discussed in the previous chapter, the components presented in this design will be addressed with respect to the future work of this research.

1. Transcribing verbalisations is a prerequisite step that has been carried out manually so far. However in order to employ the think aloud protocol as an automatic approach for assessing UX, speech to text recognition techniques are required to be investigated and accordingly integrated in this system.

2. Manual segmentation (unit of analysis) is a tedious preprocessing step similar to transcription. Alternating this task by automatic approaches could greatly facilitate think aloud data analysis. Further investigation is required in order to derive the approximate number of words per utterance. The combination of this information with objective metrics such as the silence pause could result in automatic segmentation, at the same time corresponding to the subjective criterion as "semantically meaningful units". Nevertheless, lightweight manual checking should be provided to ensure the reliability of the automatic segmentation.

3. To gain insight into users' emotional experiences, accurate data analysis of verbal data is essential. Integration of multiple measures increases the validity of data, besides revealing richer information about what has been experienced.

   Future research will be directed into the investigation on automatic analysis of the verbal and non-verbal contents, synchronisation and final fusion of the processed data.

   In this regard, ML techniques will be explored for the optimisation of the automatic analysis of verbal and non-verbal contents. For data fusion, multiple measures obtained for a given segment should be synchronised. Given analysing verbalisations based on the unit of analysis, facial expressions elicited within the timeframe of each unit (segment) are required to be aggregated and similar to the vocal expressions be normalised.

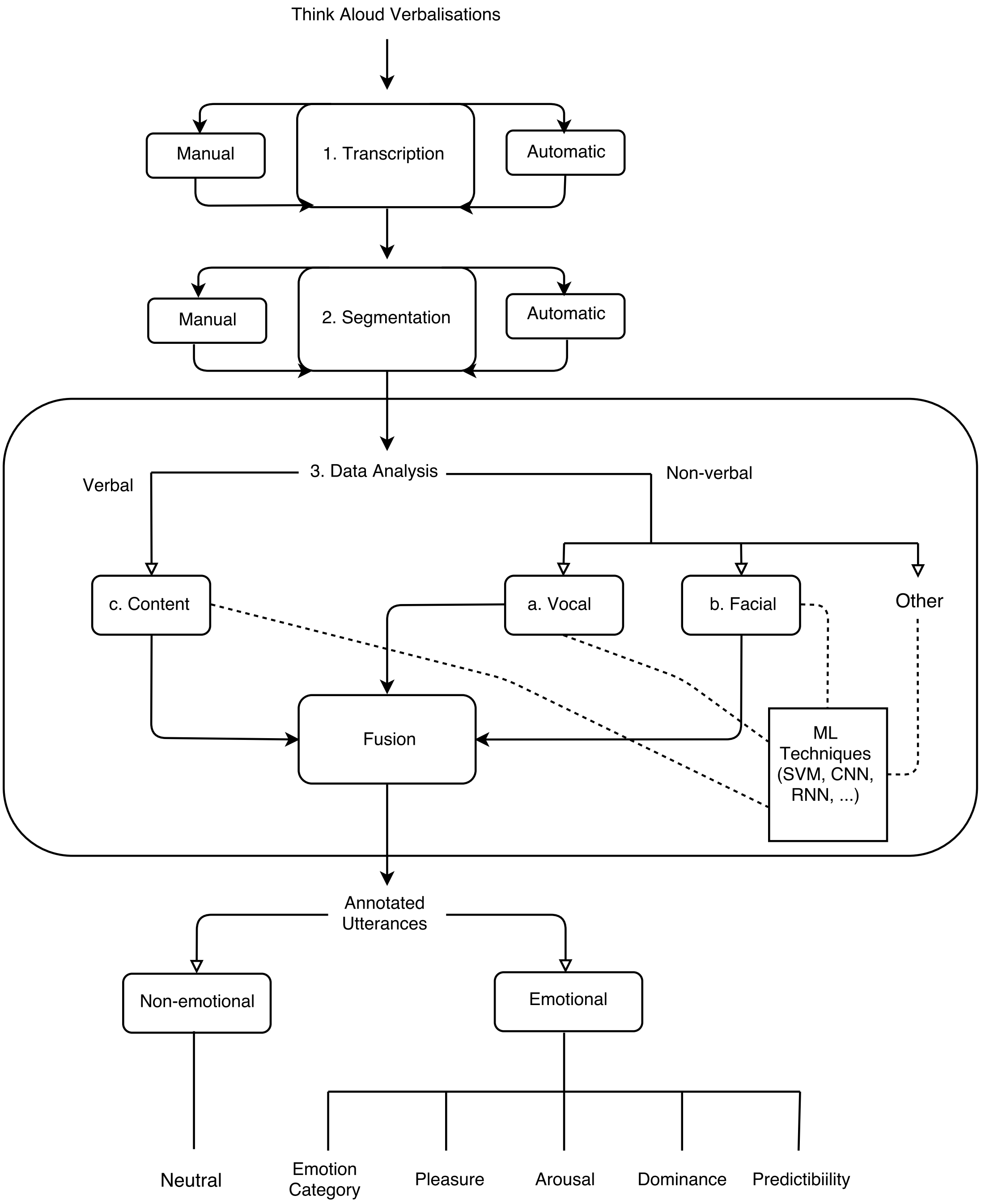


Figure 6.1: Components of automatic assessment tool based on the think aloud protocol.

> The output of this assessment tool will be annotated utterances based on categorical and dimensional labels. As it was explained before (Section 6.1.5), Predictability is a dimension that could give insight into the hedonic aspect of the experience such as perceived novelty and stimulation. Hence, this dimension will be considered for future analysis.

Overall, this emerging research area is facing a number of challenges and opportunities for future applications. Especially, advances in better understanding people's emotional experiences can contribute to supporting well-being (Section 1.1). Given the significant role of digital technology in modern life, pursuing happiness through technology is pervasive (Calvo and Peters, 2014). However, if digital technologies have not actively supported well-being, it can be due to this reason that there is yet skepticism regarding the measurability of well-being in relation to the digital products/services (Calvo and Peters, 2014).

The work presented in this thesis may be some small steps towards mitigating the uncertainty about the feasibility of measuring well-being; which hopefully will lead to bigger action in the future.

Appendix A

Questionnaire

Description

In order to complete the analysis of this research, we would like to ask you some questions regarding your experience while you were interacting with the online computer store (http://www.compdealers.com) that you visited in this study.

This survey consists of 6 sections. Please read the instructions in each section carefully. Throughout this survey "*online store"*, refers to the online computer store (**CompDealers**) that you just visited.

Please enter your name:

DES

This section consists of a number of words that describe different feelings. Please indicate the extent to which each word describes your feelings while visiting the online store [CompDealers].

Record your answers by clicking the appropriate number on the five-point scale following each word. Presented below is the scale for indicating the degree to which each word describes your feelings while you were visiting the online store.

1	2	3	4	5
Very slightly or not at all	Slightly	Moderately	Considerably	Very strongly

In deciding your answer to a given item or word, consider the feeling connoted or defined by that word. Then, if during your experience of visiting the online store you felt that way *very slightly* or *not at all*, you would click the number *1* on the scale; if you felt that way to a *moderate* degree, you would click *3*; if you felt that way very *strongly*, you would click *5*, and so forth.

It is not necessary to ponder; the first answer you decide on for a given word is probably the most valid.

	1	2	3	4	5
Attentive	○	○	○	○	○
Concentrating	○	○	○	○	○
Alert	○	○	○	○	○
Delighted	○	○	○	○	○
Happy	○	○	○	○	○
Joyful	○	○	○	○	○
Surprise	○	○	○	○	○
Amazed	○	○	○	○	○
Astonished	○	○	○	○	○
Downhearted	○	○	○	○	○
Sad	○	○	○	○	○
Discouraged	○	○	○	○	○
Enraged	○	○	○	○	○
Angry	○	○	○	○	○
Mad	○	○	○	○	○

AttrakDiff

In the table below you will find 10 pairs of contrasting attributes, where you place your choice between two attributes indicates your view about the quality of the online store that you just visited in this study [CompDealers].

For example:

Disagreeable ○ ○ ○ ○ ● ○ ○ Likable

This choice tells us that the CompDealers store is somewhat likeable, but there is still room for improvement (Note: *There is no right or wrong answer. Your personal opinion is what counts.*)

Confusing	○	○	○	○	○	○	○	Structured
Impractical	○	○	○	○	○	○	○	Practical
Unpredictable	○	○	○	○	○	○	○	Predictable
Complicated	○	○	○	○	○	○	○	Simple
Dull	○	○	○	○	○	○	○	Captivating
Tacky	○	○	○	○	○	○	○	Stylish
Cheap	○	○	○	○	○	○	○	Premium
Unimaginative	○	○	○	○	○	○	○	Creative
Bad	○	○	○	○	○	○	○	Good
Ugly	○	○	○	○	○	○	○	Beautiful

PAD

Please mark the one (out of nine) which best represents your feeling while visiting the online store [CompDealers].

(For example: in the first line, if you felt more unhappy than happy, mark the square that is closer to the word ‘‘unhappy’’, according to the extent that you felt unhappy. If you felt unhappy and happy to the same extent, mark the middle square).

Unhappy	○	○	○	○	○	○	○	○	○	Happy
Disappointed	○	○	○	○	○	○	○	○	○	Satisfied
Dispairing	○	○	○	○	○	○	○	○	○	Hopeful
Bored	○	○	○	○	○	○	○	○	○	Relaxed
Annoyed	○	○	○	○	○	○	○	○	○	Content
Languid	○	○	○	○	○	○	○	○	○	Energetic
Calm	○	○	○	○	○	○	○	○	○	Excited
Sleepy	○	○	○	○	○	○	○	○	○	Wide-awake
Slow	○	○	○	○	○	○	○	○	○	Restless
Unaroused	○	○	○	○	○	○	○	○	○	Aroused
Serene	○	○	○	○	○	○	○	○	○	Enthusiastic
Helpless	○	○	○	○	○	○	○	○	○	In control
Insignificant	○	○	○	○	○	○	○	○	○	Respected
Submissive	○	○	○	○	○	○	○	○	○	Dominant
Guided	○	○	○	○	○	○	○	○	○	Autonomous
Passive	○	○	○	○	○	○	○	○	○	Active
Restricted	○	○	○	○	○	○	○	○	○	Free

Aesthetics and Usability

Please indicate the extent to which you agree or disagree with each of the following descriptions regarding the design of the online store that you have just visited [CompDealers].

(1=strongly disagree; 7=strongly agree).

	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
Clean	○	○	○	○	○	○	○
Pleasant	○	○	○	○	○	○	○
Symmetrical	○	○	○	○	○	○	○
Aesthetic	○	○	○	○	○	○	○
Original	○	○	○	○	○	○	○
Sophisticated	○	○	○	○	○	○	○
Spectacular	○	○	○	○	○	○	○
Creative	○	○	○	○	○	○	○
It was convenient navigating in CompDealers	○	○	○	○	○	○	○
The buying process in CompDealers is simple	○	○	○	○	○	○	○
It's easy using CompDealers	○	○	○	○	○	○	○
One can find information easily in CompDealers	○	○	○	○	○	○	○

Please indicate the extent to which you agree or disagree with each of the following descriptions regarding the online store [CompDealers].

(1= strongly disagree; 7=strongly agree).

	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
I liked the online store.	○	○	○	○	○	○	○
I enjoyed being in the online store.	○	○	○	○	○	○	○
I would like to return to the online store.	○	○	○	○	○	○	○
I will recommend to others to browse the online store.	○	○	○	○	○	○	○
The online store made me feel like buying.	○	○	○	○	○	○	○

Bibliography

Abdelwahab, M. and Busso, C. (2015). Supervised domain adaptation for emotion recognition from speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5058–5062. IEEE.

Adamczyk, P. D. and Bailey, B. P. (2004). If not now, when?: the effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 271–278. ACM.

Ahlberg, J. (2001). Candide-3-an updated parameterised face.

Alabbasi, H. A., Moldoveanu, P., and Moldoveanu, A. (2015). Real time facial emotion recognition using kinect v2 sensor. *IOSR J. Comput. Eng. Ver. II*, 17(3):2278–2661.

Albert, W. and Tullis, T. (2013). *Measuring the user experience: collecting, analyzing, and presenting usability metrics.* Newnes.

Allam, A. and Dahlan, H. M. (2013). User experience: Challenges and opportunities. *Journal of Information Systems Research and Innovation*, 3:28–36.

Aly, S., Trubanova, A., Abbott, L., White, S., and Youssef, A. (2015). Vt-kfer: A kinect-based rgbd+ time dataset for spontaneous and non-spontaneous facial expression recognition. In *Biometrics (ICB), 2015 International Conference on*, pages 90–97. IEEE.

Anckar, B. and Walden, P. (2001). Self-booking of high-and low-complexity travel products: exploratory findings. *Information Technology & Tourism*, 4(3-1):151–165.

Arnold, M. J. and Reynolds, K. E. (2003). Hedonic shopping motivations. *Journal of retailing*, 79(2):77–95.

Babin, B. J., Darden, W. R., and Griffin, M. (1994). Work and/or fun: measuring hedonic and utilitarian shopping value. *Journal of consumer research*, 20(4):644–656.

Bahreini, K., Nadolski, R., and Westera, W. (2016). Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments*, 24(3):590–605.

Baloglu, S. and Brinberg, D. (1997). Affective images of tourism destinations. *Journal of travel research*, 35(4):11–15.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614.

Bänziger, T., Pirker, H., and Scherer, K. (2006). Gemep-geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions. In *Proceedings of LREC*, volume 6, pages 15–019.

Bänziger, T. and Scherer, K. R. (2007). Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus. In *International conference on affective computing and intelligent interaction*, pages 476–487. Springer.

Bargas-Avila, J. A. and Hornbæk, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2689–2698. ACM.

Barrett, L. F. and Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current directions in psychological science*, 8(1):10–14.

Batliner, A., Steidl, S., and Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus. In *Proc. of a Satellite Workshop of LREC*, pages 28–31.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., et al. (2011). Whodunnit–searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language*, 25(1):4–28.

Battarbee, K. and Koskinen, I. (2005). Co-experience: user experience as interaction. *CoDesign*, 1(1):5–18.

Bedoya-Jaramillo, S., Orozco-Arroyave, J., Arias-Londoño, J., and Vargas-Bonilla, J. (2013). Emotion recognition from telephone speech using acoustic and nonlinear features. In *Security Technology (ICCST), 2013 47th International Carnahan Conference on*, pages 1–5. IEEE.

Bennett, C. and Rudnicky, A. I. (2002). The carnegie mellon communicator corpus.

Bergstrom, J. R. and Schall, A. (2014). *Eye tracking in user experience design*. Elsevier.

Boersma, P. P. G. et al. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5.

Boren, T. and Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication*, 43(3):261–278.

Bowers, V. A. and Snyder, H. L. (1990). Concurrent versus retrospective verbal protocol for comparing window usability. In *Proceedings of the Human Factors Society Annual Meeting*, volume 34, pages 1270–1274. SAGE Publications Sage CA: Los Angeles, CA.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Bruun, A. and Ahm, S. (2015). Mind the gap! comparing retrospective and concurrent ratings of emotion in user experience evaluation. In *Human-Computer Interaction*, pages 237–254. Springer.

Bruun, A., Law, E. L.-C., Heintz, M., and Alkly, L. H. (2016a). Understanding the relationship between frustration and the severity of usability problems: What can psychophysiological data (not) tell us? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3975–3987. ACM.

Bruun, A., Law, E. L.-C., Heintz, M., and Eriksen, P. S. (2016b). Asserting real-time emotions through cued-recall: Is it valid? In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, page 37. ACM.

Bui, M. and Kemp, E. (2013). E-tail emotion regulation: examining online hedonic product purchases. *International Journal of Retail & Distribution Management*, 41(2):155–170.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.

Burr, B. (2006). Vaca: a tool for qualitative video analysis. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 622–627. ACM.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Busso, C., Bulut, M., Narayanan, S., Gratch, J., and Marsella, S. (2013). Toward effective automatic recognition systems of emotion in speech. *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, pages 110–127.

Busso, C. and Narayanan, S. (2008a). Recording audio-visual emotional databases from actors: a closer look. In *Second international workshop on emotion: corpora for research on emotion and affect, international conference on language resources and evaluation (LREC 2008)*, pages 17–22.

Busso, C. and Narayanan, S. S. (2008b). The expression and perception of emotions: comparing assessments of self versus others. In *Interspeech*, pages 257–260.

Calvo, R. A. and D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37.

Calvo, R. A. and Peters, D. (2014). *Positive computing: technology for wellbeing and human potential.* MIT Press.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological bulletin*, 54(4):297.

Caruana, A. and Ewing, M. T. (2010). How corporate reputation, quality, and value influence online loyalty. *Journal of Business Research*, 63(9):1103–1110.

Chandra, A. and Calderon, T. (2005). Challenges and constraints to the diffusion of biometrics in information systems. *Communications of the ACM*, 48(12):101–106.

Cheng, F.-F., Wu, C.-S., and Yen, D. C. (2009). The effect of online store atmosphere on consumer's emotional responses–an experimental study of music and colour. *Behaviour & Information Technology*, 28(4):323–334.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The journal of the learning sciences*, 6(3):271–315.

Clore, G. L., Schwarz, N., and Conway, M. (1994). Affective causes and consequences of social information processing. *Handbook of social cognition*, 1:323–417.

Cohn, J. F. (2006). Foundations of human computing: facial expression and emotion. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 233–238. ACM.

Cooke, L. (2010). Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication*, 53(3):202–215.

Cooke, L. and Cuddihy, E. (2005). Using eye tracking to address limitations in think-aloud protocol. In *Professional Communication Conference, 2005. IPCC 2005. Proceedings. International*, pages 653–658. IEEE.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Craik, F. I. and Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, 11(6):671–684.

Crowley, A. E., Spangenberg, E. R., and Hughes, K. R. (1992). Measuring the hedonic and utilitarian dimensions of attitudes toward product categories. *Marketing letters*, 3(3):239–249.

Dai, W., Han, D., Dai, Y., and Xu, D. (2015). Emotion recognition and affective computing on vocal social media. *Information & Management*, 52(7):777–788.

Darwin, C. (1998). *The expression of the emotions in man and animals*. Oxford University Press, USA.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

Desmet, P. and Hekkert, P. (2007). Framework of product experience. *International journal of design*, 1(1).

Desmet, P. M. (2012). Faces of product pleasure: 25 positive emotions in human-product interactions. *International Journal of Design, 6 (2), 2012.*

Desmet, P. M. and Pohlmeyer, A. E. (2013). Positive design: An introduction to design for subjective well-being. *International Journal of Design*, 7(3).

Ding, C. G. and Lin, C.-H. (2012). How does background music tempo work for online shopping? *Electronic Commerce Research and Applications*, 11(3):299–307.

DIS, I. (2009). 9241-210: 2010. ergonomics of human system interaction-part 210: Human-centred design for interactive systems. *International Standardization Organization (ISO). Switzerland.*

D'mello, S. and Graesser, A. (2007). Mind and body: Dialogue and posture for affect detection in learning environments. *Frontiers in Artificial Intelligence and Applications*, 158:161.

D'Mello, S., Lehman, B., Pekrun, R., and Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech communication*, 40(1):33–60.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., et al. (2007). The humaine database: addressing the collection and annotation of naturalistic and induced emotional data. *Affective computing and intelligent interaction*, pages 488–500.

Douglas-Cowie, E. and members, W. (2004). Preliminary plans for exemplars: Databases. `http://emotion-research.net/projects/humaine/deliverables/D5c.pdf`.

Dumas, J. S. and Redish, J. (1999). *A practical guide to usability testing.* Intellect books.

Ekkekakis, P. (2013). *The measurement of affect, mood, and emotion: A guide for health-behavioral research.* Cambridge University Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Ekman, P. (1999). Facial expressions. *Handbook of cognition and emotion*, 16:301–320.

Ekman, P. and Friesen, W. V. (1977). Facial action coding system.

Ekman, P., Friesen, W. V., and Hager, J. (1978). The facial action coding system (facs): A technique for the measurement of facial action. palo alto. *CA: Consulting Psychologists Press, Inc. Ekman, P. Levenson. RW, & Friesen WV (1983). Autonomic nervous system activity distinguishes among emotions. Science*, 221:1208–12.

Ekman, P. and Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press, USA.

Ekman, P. and Scherer, K. (1984). Expression and the nature of emotion. *Approaches to emotion*, 3:19–344.

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

Ellsworth, P. C. and Scherer, K. R. (2003). Appraisal processes in emotion. *Handbook of affective sciences*, 572:V595.

Engberg, I. S., Hansen, A. V., Andersen, O., and Dalsgaard, P. (1997). Design, recording and verification of a danish emotional speech database. In *Fifth European Conference on Speech Communication and Technology.*

Epp, C., Lippold, M., and Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 715–724. ACM.

Ericsson, K. A. and Simon, H. A. (1993). *Protocol analysis.* MIT press Cambridge, MA.

Eroglu, S. A., Machleit, K. A., and Davis, L. M. (2001). Atmospheric qualities of online retailing: A conceptual model and implications. *Journal of Business research*, 54(2):177–184.

Eyben, F. (2015). *Real-time speech and music classification by large audio feature space extraction.* Springer.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Eyben, F., Wöllmer, M., and 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*, pages 1–6. IEEE.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Feldman Barrett, L. and Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of personality and social psychology*, 74(4):967.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Floh, A. and Madlberger, M. (2013). The role of atmospheric cues in online impulse-buying behavior. *Electronic Commerce Research and Applications*, 12(6):425–439.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.

Forlizzi, J. and Battarbee, K. (2004). Understanding experience in interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 261–268. ACM.

Forlizzi, J. and Ford, S. (2000). The building blocks of experience: an early framework for interaction designers. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 419–423. ACM.

Forrester (2017). European online retail forecast. `https://go.forrester.com/`.

Fredrickson, B. L. and Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of personality and social psychology*, 65(1):45.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92.

Frijda, N. H. and Philipszoon, E. (1963). Dimensions of recognition of expression. *The Journal of Abnormal and Social Psychology*, 66(1):45.

Gentile, C., Spiller, N., and Noci, G. (2007). How to sustain the customer experience:: An overview of experience components that co-create value with the customer. *European Management Journal*, 25(5):395–410.

Goleman, D. (2006). *Emotional intelligence.* Bantam.

Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., and Lester, J. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*.

Gray, W. D. and Salzman, M. C. (1998). Damaged merchandise? a review of experiments that compare usability evaluation methods. *Human–computer interaction*, 13(3):203–261.

Grimm, M., Kroschel, K., Mower, E., and Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10):787–800.

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

Gross, A. and Bongartz, S. (2012). Why do i like it?: investigating the product-specificity of user experience. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, pages 322–330. ACM.

Ha, H.-Y. and Perks, H. (2005). Effects of consumer perceptions of brand experience on the web: Brand familiarity, satisfaction and brand trust. *Journal of Consumer Behaviour*, 4(6):438–452.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hand, D. J. (2004). Measurement theory and practice.

Hansen, J. H. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech communication*, 20(1-2):151–173.

Haq, S., Jackson, P. J., and Edge, J. (2009). Speaker-dependent audio-visual emotion recognition. In *AVSP*, pages 53–58.

Hartmann, J., Sutcliffe, A., and De Angeli, A. (2007). Investigating attractiveness in web user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 387–396. ACM.

Hassenzahl, M. (2003). The thing and i: understanding the relationship between user and product. *Funology*, pages 31–42.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-computer interaction*, 19(4):319–349.

Hassenzahl, M. (2008). User experience (ux): towards an experiential perspective on product quality. In *Proceedings of the 20th Conference on l'Interaction Homme-Machine*, pages 11–15. ACM.

Hassenzahl, M. (2013). User experience and experience design. *The Encyclopedia of Human-Computer Interaction,*.

Hassenzahl, M., Beu, A., and Burmester, M. (2001). Engineering joy. *Ieee Software*, 18(1):70–76.

Hassenzahl, M., Kekez, R., and Burmester, M. (2002). The importance of a software's pragmatic quality depends on usage modes. In *Proceedings of the 6th international conference on Work With Display Units (WWDU 2002)*, pages 275–276. ERGONOMIC Institut für Arbeits-und Sozialforschung Berlin.

Hassenzahl, M. and Monk, A. (2010). The inference of perceived usability from beauty. *Human–Computer Interaction*, 25(3):235–260.

Hassenzahl, M., Schöbel, M., and Trautmann, T. (2008). How motivational orientation influences the evaluation and choice of hedonic and pragmatic interactive products: The role of regulatory focus. *Interacting with Computers*, 20(4-5):473–479.

Hassenzahl, M. and Ullrich, D. (2007). To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. *Interacting with computers*, 19(4):429–437.

Hazlett, R. L. (2006). Measuring emotional valence during interactive experiences: boys at video game play. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1023–1026. ACM.

Hertzum, M., Borlund, P., and Kristoffersen, K. B. (2015). What do thinking-aloud participants say? a comparison of moderated and unmoderated usability sessions. *International Journal of Human-Computer Interaction*, 31(9):557–570.

Hertzum, M., Hansen, K. D., and Andersen, H. H. (2009). Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, 28(2):165–181.

Hertzum, M. and Holmegaard, K. D. (2013). Thinking aloud in the presence of interruptions and time constraints. *International Journal of Human-Computer Interaction*, 29(5):351–364.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Holbrook, M. B. and Hirschman, E. C. (1982). The experiential aspects of consumption: Consumer fantasies, feelings, and fun. *Journal of consumer research*, 9(2):132–140.

Hsieh, J.-K., Hsieh, Y.-C., Chiu, H.-C., and Yang, Y.-R. (2014). Customer response to web site atmospherics: Task-relevant cues, situational involvement and pad. *Journal of Interactive Marketing*, 28(3):225–236.

Izard, C. E. (2013). *Human emotions.* Springer Science & Business Media.

Jones, C. and Deeming, A. (2008). Affective human-robotic interaction. In *Affect and emotion in human-computer interaction*, pages 175–185. Springer.

Jordan, P. W. (2002). *Designing pleasurable products: An introduction to the new human factors.* CRC press.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Kahneman, D. et al. (1999). Objective happiness. *Well-being: The foundations of hedonic psychology*, 3:25.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., and Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological science*, 4(6):401–405.

Kahneman, D., Kahneman, D., Tversky, A., et al. (2003). Experienced utility and objective happiness: A moment-based approach. *The psychology of economic decisions*, 1:187–208.

Kahneman, D., Wakker, P. P., and Sarin, R. (1997). Back to bentham? explorations of experienced utility. *The quarterly journal of economics*, 112(2):375–406.

Kaltcheva, V. D. and Weitz, B. A. (2006). When should a retailer create an exciting store environment? *Journal of marketing*, 70(1):107–118.

Kaplan, K. J. (1972). On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological bulletin*, 77(5):361.

Kawas, S., Karalis, G., Wen, T., and Ladner, R. E. (2016). Improving real-time captioning experiences for deaf and hard of hearing students. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 15–23. ACM.

Kim, H.-C., Pang, S., Je, H.-M., Kim, D., and Bang, S. Y. (2003). Constructing support vector machine ensemble. *Pattern recognition*, 36(12):2757–2767.

Kleinginna, P. R. and Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379.

Koolagudi, S. G. and Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117.

Kouklia, C. and Audibert, N. (2013). Expressivity conveyed by contextualized vs. non-contextualized ?neutral? acted speech: which control condition for expressive speech modeling? In *First International Workhop on Affective Social Speech Signals*.

Krahmer, E. and Ummelen, N. (2004). Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication*, 47(2):105–117.

Kudoh, T. and Matsumoto, Y. (2000). Use of support vector learning for chunk identification. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 142–144. Association for Computational Linguistics.

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., and Sinnelä, A. (2011). Ux curve: A method for evaluating long-term user experience. *Interacting with Computers*, 23(5):473–483.

Kujala, S., Walsh, T., Nurkka, P., and Crisan, M. (2014). Sentence completion for understanding users and evaluating user experience. *Interacting with Computers*, 26(3):238–255.

Kuncheva, L. I. and Rodríguez, J. J. (2014). A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems*, 38(2):259–275.

Kwon, O.-W., Chan, K., Hao, J., and Lee, T.-W. (2003). Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology*.

Larson, R. and Csikszentmihalyi, M. (1983). The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*.

Lasecki, W. S., Miller, C. D., Naim, I., Kushalnagar, R., Sadilek, A., Gildea, D., and Bigham, J. P. (2017). Scribe: Deep integration of human and machine intelligence to caption speech in real time. *COMMUNICATIONS OF THE ACM*, 60(11).

Lavie, T. and Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International journal of human-computer studies*, 60(3):269–298.

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., and Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 719–728. ACM.

Law, E. L.-C. and van Schaik, P. (2010). Modelling user experience–an agenda for research and practice. *Interacting with computers*, 22(5):313–322.

Law, E. L.-C., van Schaik, P., and Roto, V. (2014). Attitudes towards user experience (ux) measurement. *International Journal of Human-Computer Studies*, 72(6):526–541.

Lee, C.-C., Mower, E., Busso, C., Lee, S., and Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9):1162–1171.

Lee, S., Yildirim, S., Kazemzadeh, A., and Narayanan, S. (2005). An articulatory study of emotional speech production. In *Interspeech*, pages 497–500.

Lefter, I., Rothkrantz, L., Wiggers, P., and van Leeuwen, D. (2010). Emotion recognition from speech by combining databases and fusion of classifiers. In *Text, Speech and Dialogue*, pages 353–360. Springer.

Li, Y., Tao, J., Chao, L., Bao, W., and Liu, Y. (2017). Cheavd: a chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, 8(6):913–924.

Liberman, M. (2002). Emotional prosody speech and transcripts ldc2002s28. `https://catalog.ldc.upenn.edu/LDC2002S28`.

Lin, A., Gregor, S., and Ewing, M. (2008). Developing a scale to measure the enjoyment of web experiences. *Journal of Interactive Marketing*, 22(4):40–57.

Littlewort, G. C., Bartlett, M. S., Salamanca, L. P., and Reilly, J. (2011). Automated measurement of children's facial expressions during problem solving tasks. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 30–35. IEEE.

Liu, D., Bias, R. G., Lease, M., and Kuipers, R. (2012). Crowdsourcing for usability testing. *Proceedings of the Association for Information Science and Technology*, 49(1):1–10.

Lopatovska, I. and Arapakis, I. (2011). Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. *Information Processing & Management*, 47(4):575–592.

Ma, J., Jin, H., Yang, L. T., and Tsai, J. J.-P. (2006). Ubiquitous intelligence and computing: Third international conference, uic 2006, wuhan, china, september 3-6, 2006. *Proceedings (Lecture Notes in Computer Science), Springer-Verlag New York, Inc., Secaucus, NJ.*

Mahlke, S. and Minge, M. (2008). Consideration of multiple components of emotions in human-technology interaction. In *Affect and emotion in human-computer interaction*, pages 51–62. Springer.

Mahlke, S. and Thüring, M. (2007). Studying antecedents of emotional experiences in interactive contexts. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 915–918. ACM.

Makri, S., Blandford, A., and Cox, A. L. (2010). This is what i'm doing and why: reflections on a think-aloud study of dl users' information behaviour. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 349–352. ACM.

Mandryk, R. L., Inkpen, K. M., and Calvert, T. W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & information technology*, 25(2):141–158.

Mao, Q.-r., Pan, X.-y., Zhan, Y.-z., and Shen, X.-j. (2015). Usingkinect for real-time emotion recognition via facial expressions. *Frontiers of Information Technology & Electronic Engineering*, 16(4):272–282.

Marchi, E., Eyben, F., Hagerer, G., and Schuller, B. W. (2016). Real-time tracking of speakers' emotions, states, and traits on mobile platforms. In *INTERSPEECH*, pages 1182–1183.

Marchi, E., Schuller, B., Baron-Cohen, S., Golan, O., Bölte, S., Arora, P., and Häb-Umbach, R. (2015a). Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages. In *Sixteenth Annual Conference of the International Speech Communication Association.*

Marchi, E., Schuller, B., Baron-Cohen, S., Lassalle, A., O?Reilly, H., Pigat, D., Golan, O., Friedenson, S., Tal, S., Bölte, S., et al. (2015b). Voice emotion games: Language and emotion in the voice of children with autism spectrum condition.

In *Proc. of the 3rd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2015) as part of the 20th ACM International Conference on Intelligent User Interfaces, IUI.*

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The enterface?05 audio-visual emotion database. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 8–8. IEEE.

Mazaheri, E., Richard, M.-O., and Laroche, M. (2012). The role of emotions in online consumer behavior: a comparison of search, experience, and credence services. *Journal of Services Marketing*, 26(7):535–550.

McCarthy, J. and Wright, P. (2004). Technology as experience. *interactions*, 11(5):42–43.

McDonald, S., Zhao, T., and Edwards, H. M. (2013). Dual verbal elicitation: the complementary use of concurrent and retrospective reporting within a usability test. *International Journal of Human-Computer Interaction*, 29(10):647–660.

Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs.*

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Mehrabian, A. and Russell, J. A. (1974). *An approach to environmental psychology.* the MIT Press.

Mehu, M. and Scherer, K. R. (2015). Emotion categories and dimensions in the facial communication of affect: An integrated approach. *Emotion*, 15(6):798.

Microsoft (2015a). Kinect for windows sdk 2.0. `https://msdn.microsoft.com/en-us/library/dn785525.aspx`.

Microsoft (2015b). Kinect hardware. `https://dev.windows.com/en-us/kinect/hardware`.

Miron-Shatz, T., Stone, A., and Kahneman, D. (2009). Memories of yesterday?s emotions: Does the valence of experience affect the memory-experience gap? *Emotion*, 9(6):885.

Morgan-Thomas, A. and Veloutsou, C. (2013). Beyond technology acceptance: Brand relationships and online brand experience. *Journal of Business Research*, 66(1):21–27.

Morrison, D., Wang, R., and De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech communication*, 49(2):98–112.

Moutinho, L. (1987). Consumer behaviour in tourism. *European journal of marketing*, 21(10):5–44.

Nielsen, J. (1994a). Estimating the number of subjects needed for a thinking aloud test. *International journal of human-computer studies*, 41(3):385–397.

Nielsen, J. (1994b). *Usability engineering*. Elsevier.

Nielsen, J., Clemmensen, T., and Yssing, C. (2002). Getting access to what goes on in people's heads?: reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction*, pages 101–110. ACM.

Niforatos, E., Karapanos, E., Langheinrich, M., Wurhofer, D., Krischkowsky, A., Obrist, M., and Tscheligi, M. (2015). emotion: retrospective in-car user experience evaluation. In *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 118–123. ACM.

Nowak, M., Kim, J., Kim, N. W., and Nass, C. (2012). Social visualization and negotiation: effects of feedback configuration and status. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1081–1090. ACM.

O'Brien, H. L. (2010). The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping experiences. *Interacting with computers*, 22(5):344–352.

O'Brien, H. L. and Lebow, M. (2013). Mixed-methods approach to measuring user experience in online news interactions. *Journal of the Association for Information Science and Technology*, 64(8):1543–1556.

Okazaki, S. (2006). Excitement or sophistication? a preliminary exploration of online brand personality. *International Marketing Review*, 23(3):279–303.

Olsson, T., Lagerstam, E., Kärkkäinen, T., and Väänänen-Vainio-Mattila, K. (2013). Expected user experience of mobile augmented reality services: a user study in the context of shopping centres. *Personal and ubiquitous computing*, 17(2):287–304.

Oxley, B. (2005). Hostgator. `http://www.hostgator.com`.

Pantic, M. and Rothkrantz, L. J. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390.

Pantic, M. and Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1424–1445.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Peterson, R. A., Balasubramanian, S., and Bronnenberg, B. J. (1997). Exploring the implications of the internet for consumer marketing. *Journal of the Academy of Marketing science*, 25(4):329–346.

Petridis, S., Martinez, B., and Pantic, M. (2013). The mahnob laughter database. *Image and Vision Computing*, 31(2):186–202.

Petrie, H. and Precious, J. (2010). Measuring user experience of websites: Think aloud protocols and an emotion word prompt list. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 3673–3678. ACM.

Pfister, T. and Robinson, P. (2011). Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *IEEE Transactions on Affective Computing*, 2(2):66–78.

Phillips, D. L. and Clancy, K. J. (1972). Some effects of social desirability in survey studies. *American Journal of Sociology*, 77(5):921–940.

Picard, R. W. and Picard, R. (1997). *Affective computing*, volume 252. MIT press Cambridge.

Pittam, J. and Scherer, K. R. (1993). *Vocal expression and communication of emotion.* Guilford Press.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis.* Harpercollins College Division.

Porat, T. and Tractinsky, N. (2012). It's a pleasure buying here: The effects of web-store design on consumers' emotions and attitudes. *Human–Computer Interaction*, 27(3):235–276.

Pressley, M. and Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* Routledge.

Qualtrics (2000). Qualtrics. `http://www.qualtrics.com`.

Rahman, T. and Busso, C. (2012). A personalized emotion recognition system using an unsupervised feature adaptation scheme. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5117–5120. IEEE.

Raikwal, J. and Saxena, K. (2012). Performance evaluation of svm and k-nearest neighbor algorithm over medical data set. *International Journal of Computer Applications*, 50(14).

Rajoo, R. and Aun, C. C. (2016). Influences of languages in speech emotion recognition: A comparative study using malay, english and mandarin languages. In *Computer Applications & Industrial Electronics (ISCAIE), 2016 IEEE Symposium on*, pages 35–39. IEEE.

Rao, K. S. and Yegnanarayana, B. (2006). Prosody modification using instants of significant excitation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):972–980.

Redelmeier, D. A. and Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, 66(1):3–8.

Rezende-Parker, A. M., Morrison, A. M., and Ismail, J. A. (2003). Dazed and confused? an exploratory study of the image of brazil as a travel destination. *Journal of Vacation Marketing*, 9(3):243–259.

Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., Lalanne, D., and Schuller, B. (2015). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30.

Ringeval, F., Marchi, E., Grossard, C., Xavier, J., Chetouani, M., Cohen, D., and Schuller, B. (2016). Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. In *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, pages 1210–1214.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE.

Robson, C. (2002). Real world research: a resource for social scientists and practitioner. *Adapting Open Innovation in ICT Ecosystem Dynamics References Real World Research: A Resource for Social Scientists and Practitioner*, page 270.

Rosenberg, A. (2012). Classifying skewed data: Importance weighting to optimize average recall. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Rubin, J. and Chisnell, D. (2008). *Handbook of usability testing: howto plan, design, and conduct effective tests*. John Wiley & Sons.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Russell, J. A., Lewicka, M., and Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, 57(5):848–856.

Russell, J. A. and Mehrabian, A. (1974). Distinguishing anger and anxiety in terms of emotional response factors. *Journal of consulting and clinical psychology*, 42(1):79.

Safavi, R. (2009). Interface design issues to enhance usability of e-commerce websites and systems. In *Computer Technology and Development, 2009. ICCTD'09. International Conference on*, volume 1, pages 277–281. IEEE.

Sanderson, P. M. (1990). Verbal protocol analysis in three experimental domains using shapa. In *Proceedings of the Human Factors Society Annual Meeting*, volume 34, pages 1280–1284. SAGE Publications Sage CA: Los Angeles, CA.

Sato, N. and Obuchi, Y. (2007). Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, 2(3):835–848.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):143.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1):227–256.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Scherer, K. R., Schorr, A., and Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research.* Oxford University Press.

Schmettow, M. (2012). Sample size in usability studies. *Communications of the ACM*, 55(4):64–70.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., et al. (2007a). The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *Eighth Annual Conference of the International Speech Communication Association.*

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

Schuller, B., Müeller, R., Höernler, B., Höethker, A., Konosu, H., and Rigoll, G. (2007b). Audiovisual recognition of spontaneous interest within conversations. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 30–37. ACM.

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., and Konosu, H. (2009a). Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12):1760–1774.

Schuller, B., Rigoll, G., Grimm, M., Kroschel, K., Moosmayr, T., and Ruske, G. (2007c). Effects of in-car noise-conditions on the recognition of emotion within speech. *Fortschritte der Akustik*, 33(1):305.

Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden markov model-based speech emotion recognition. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 1, pages I–401. IEEE.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., and Narayanan, S. (2013). Paralinguistics in speech and language?state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. S. (2010a). The interspeech 2010 paralinguistic challenge. In *Eleventh Annual Conference of the International Speech Communication Association.*

Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J. (2011b). The interspeech 2011 speaker state challenge. In *Twelfth Annual Conference of the International Speech Communication Association*.

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

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

Schuller, B., Zhang, Z., Weninger, F., and Rigoll, G. (2011c). Using multiple databases for training in emotion recognition: To unite or to vote? In *Twelfth Annual Conference of the International Speech Communication Association*.

Schuller, B. W., Steidl, S., Batliner, A., et al. (2009c). The interspeech 2009 emotion challenge. In *Interspeech*, volume 2009, pages 312–315.

Schuller, B. W., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., Van Son, R., Weninger, F., Eyben, F., Bocklet, T., et al. (2012). The interspeech 2012 speaker trait challenge. In *Interspeech*, volume 2012, pages 254–257.

Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition & Emotion*, 14(4):433–440.

Shadiev, R., Hwang, W.-Y., Chen, N.-S., and Yueh-Min, H. (2014). Review of speech-to-text recognition technology for enhancing learning. *Journal of Educational Technology & Society*, 17(4):65.

Shami, M. and Verhelst, W. (2007). Automatic classification of expressiveness in speech: a multi-corpus study. In *Speaker classification II*, pages 43–56. Springer.

Sherwani, D. and Stumpf, S. (2014). Toward helping users in assessing the trustworthiness of user-generated reviews. In *Proceedings of the 28th International BCS Human Computer Interaction Conference on HCI 2014-Sand, Sea and Sky-Holiday HCI*, pages 120–129. BCS.

Soleimani, S. and Law, E. L.-C. (2015). The influence of motivation on emotional experience in e-commerce. In *Human-Computer Interaction*, pages 281–288. Springer.

Soleimani, S. and Law, E. L.-C. (2017). What can self-reports and acoustic data analyses on emotions tell us? In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pages 489–501. ACM.

Solera-Urena, R., Moniz, H., Batista, F., Cabarrao, V., Pompili, A., Fernández-Astudillo, R., Campos, J., Paiva, A., and Trancoso, I. (2017). A semi-supervised learning approach for acoustic-prosodic personality perception in under-resourced domains. *Proc. Interspeech 2017*, pages 929–933.

Soleymani, M., Lichtenauer, J., Pun, T., and Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55.

Staiano, J., Menéndez, M., Battocchi, A., De Angeli, A., and Sebe, N. (2012). Ux_mate: from facial expressions to ux evaluation. In *Proceedings of the Designing Interactive Systems Conference*, pages 741–750. ACM.

Stickel, C., Ebner, M., Steinbach-Nordmann, S., Searle, G., and Holzinger, A. (2009). Emotion detection: application of the valence arousal space for rapid biological usability testing to enhance universal access. *Universal Access in Human-Computer Interaction. Addressing Diversity*, pages 615–624.

Sward, D. and Macarthur, G. (2007). Making user experience a business strategy. In *E. Law et al.(eds.), Proceedings of the Workshop on Towards a UX Manifesto*, volume 3, pages 35–40.

Trickett, S. and Trafton, J. G. (2009). A primer on verbal protocol analysis. *The PSI handbook of virtual environments for training and education*, 1:332–346.

Tuch, A. N., Roth, S. P., HornbæK, K., Opwis, K., and Bargas-Avila, J. A. (2012). Is beautiful really usable? toward understanding the relation between usability, aesthetics, and affect in hci. *Computers in Human Behavior*, 28(5):1596–1607.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM.

Van Den Haak, M., De Jong, M., and Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & information technology*, 22(5):339–351.

Vapnik, V. (2013). *The nature of statistical learning theory.* Springer science & business media.

Vasalou, A. and Bänziger, T. (2006). Using the *think aloud* for affective evaluations in the lab: a work-in-progress.

Vaudable, C., Rollet, N., and Devillers, L. (2010). Annotation of affective interaction in real-life dialogs collected in a call-center. In *The Workshop Programme*, page 47.

Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181.

Vogt, T., André, E., and Bee, N. (2008). Emovoice?a framework for online recognition of emotions from voice. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 188–199. Springer.

Ward, R. D. and Marsden, P. H. (2003). Physiological responses to different web page designs. *International Journal of Human-Computer Studies*, 59(1):199–212.

Wastell, D. G. and Newman, M. (1996). Stress, control and computer system design: a psychophysiological field study. *Behaviour & Information Technology*, 15(3):183–192.

Wetherell, M. (2012). *Affect and emotion: A new social science understanding*. Sage Publications.

Williams, C. E. and Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250.

Wixon, D. (2011). Measuring fun, trust, confidence, and other ethereal constructs: it isn't that hard. *interactions*, 18(6):74–77.

Wordpress (2002). Wordpress. `https://wordpress.com`.

Wu, C.-H., Lin, J.-C., and Wei, W.-L. (2013). Two-level hierarchical alignment for semi-coupled hmm-based audiovisual emotion recognition with temporal course. *IEEE Transactions on Multimedia*, 15(8):1880–1895.

Wu, C.-H., Lin, J.-C., and Wei, W.-L. (2014). Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing*, 3:e12.

Wu, X. and Srihari, R. (2004). Incorporating prior knowledge with weighted margin support vector machines. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 326–333. ACM.

Yang, J., Yan, R., and Hauptmann, A. G. (2007). Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 188–197. ACM.

Yik, M. S., Russell, J. A., and Barrett, L. F. (1999). Structure of self-reported current affect: Integration and beyond. *Journal of personality and social psychology*, 77(3):600.

Zaman, B. and Shrimpton-Smith, T. (2006). The facereader: Measuring instant fun of use. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, pages 457–460. ACM.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58.

Zhang, B., Essl, G., and Mower Provost, E. (2016). Automatic recognition of self-reported and perceived emotion: does joint modeling help? In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 217–224. ACM.

Zhao, T., McDonald, S., and Edwards, H. M. (2014). The impact of two different think-aloud instructions in a usability test: a case of just following orders? *Behaviour & Information Technology*, 33(2):163–183.

Zimmermann, P., Gomez, P., Danuser, B., and Schär, S. (2006). Extending usability: putting affect into the user-experience. *Proceedings of NordiCHI'06*, pages 27–32.