2014-6

The Perception of Emotion from Acoustic Cues in Natural Speech

John Snel
Technological University Dublin

The Perception of Emotion from Acoustic Cues in Natural Speech

John Snel

A thesis presented to the Dublin Institute of Technology,

Digital Media Centre (DMC)

for the degree of Doctor of Philosophy

Research Supervisors:

Dr. Charlie Cullen, Dr. Sarah Jane Delany, Prof. Nick Campbell

June 2014

Abstract

Knowledge of human perception of emotional speech is imperative for the development of emotion in speech recognition systems and emotional speech synthesis. Owing to the fact that there is a growing trend towards research on spontaneous, real-life data, the aim of the present thesis is to examine human perception of emotion in naturalistic speech. Although there are many available emotional speech corpora, most contain simulated expressions. Therefore, there remains a compelling need to obtain naturalistic speech corpora that are appropriate and freely available for research. In that regard, our initial aim was to acquire suitable naturalistic material and examine its emotional content based on listener perceptions. A web-based listening tool was developed to accumulate ratings based on large-scale listening groups. The emotional content present in the speech material was demonstrated by performing perception tests on conveyed levels of Activation and Evaluation. As a result, labels were determined that signified the emotional content, and thus contribute to the construction of a naturalistic emotional speech corpus.

In line with the literature, the ratings obtained from the perception tests suggested that Evaluation (or hedonic valence) is not identified as reliably as Activation is. Emotional valence can be conveyed through both semantic and prosodic information, for which the meaning of one may serve to facilitate, modify, or conflict with the meaning of the other—particularly with naturalistic speech. The subsequent experiments aimed to investigate this concept by comparing ratings from perception tests of non-verbal speech with verbal speech. The method used

to render non-verbal speech was low-pass filtering, and for this, suitable filtering conditions were determined by carrying out preliminary perception tests. The results suggested that non-verbal naturalistic speech provides sufficiently discernible levels of Activation and Evaluation. It appears that the perception of Activation and Evaluation is affected by low-pass filtering, but that the effect is relatively small. Moreover, the results suggest that there is a similar trend in agreement levels between verbal and non-verbal speech. To date it still remains difficult to determine unique acoustical patterns for hedonic valence of emotion, which may be due to inadequate labels or the incorrect selection of acoustic parameters. This study has implications for the labelling of emotional speech data and the determination of salient acoustic correlates of emotion.

Declaration

I certify that this thesis which I now submit for examination for the award of Doctor of Philosophy, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for postgraduate study by research of the Dublin Institute of Technology and has not been submitted in whole or in part for another award in any other third level institution.

The work reported on in this thesis conforms to the principles and requirements of the DIT's guidelines for ethics in research.

DIT has permission to keep, lend or copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature_____ **Date**_____

Acknowledgements

I would like to express my gratitude to several people who have supported me in many ways during this research. First of all, thanks to my principal supervisor, Dr. Charlie Cullen, for the invaluable academic guidance and occasional confidence counselling. The research contained within this thesis would never have been possible without his knowledge. Thanks also to my second supervisor, Dr. Sarah Jane, who also gave me plenty of sound advice, and made this project possible. Thanks to both for the insightful discussions. I would also like to express gratitude to my external supervisor Prof. Nick Campbell.

It has been a great privilege to work at the Digital Media Centre, and I would like to thank all past and present members of the group. Specifically, I would like to thank my colleagues who worked closely with this project: Dr. Alexey Tarasov, Dr. Brian Vaughan, and Anna Deegan. I'd like to extend my thanks to my other colleagues, who I shared office space with, for the discussions, assistance, advice, and amusing conversations, including John McGee, Dr. Mark Dunne, Niels Schütte, Dr. Viacheslav Filonenko, Simon Bursell and Tiedong Yang. I would also like to thank Charlie Pritchard of the DMC for his help and support during the process. I would like to thank Sharon Murray, Barbara O'Shea, and Dr. Brian O'Neill of the School of Media, who have been very kind and supportive throughout this research. I would like to extend my thanks to others from the research community, with special thanks to Tanja Bänziger, Mirja Ilves, and Stefan Steidl, for help, advice, and materials. A sincere thanks to John Clifford for doing the final proof reading.

I would also like to thank my family, especially my mother and father for the mental support, and technical advice. I am grateful for my mother's suggestion to pursue this venture, for the discussions and advice on statistical issues, and for giving me the constant moral support by understanding the task at hand. Many thanks to my father for his patientce, technical advice, and proof-reading. If it wasn't for both of them, I would not have been able to undertake nor complete this thesis. Thanks to all my friends, and to all those who took part in the experiments. Finally, I would like to thank Alice McNamara for her unconditional patience, support and love. As she would say, "Boom!! ... done".

Contents

List of Figures

List of Tables

List of publications

The following is a list of publications arisen from the work within this thesis:

- Snel, J., Cullen, C.: Judging Emotion from Low-pass Filtered Naturalistic Emotional Speech. *Affective Computing and Intelligent Interaction (ACII)*, Fifth biannual Humaine Association Conference on Affective Computing and Intelligent Interaction. Geneva, Switzerland on September 2-5, 2013. `http://arrow.dit.ie/dmccon/105/`

- Snel, J., Tarasov, A., Cullen, C., Delany, S.J.: A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpora. *The 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES^3 2012)*, 2012. `http://arrow.dit.ie/dmccon/97/`

- Snel, J., Cullen, C.: Obtaining speech assets for judgement analysis on low-pass filtered emotional speech. *EmoSPACE 2011 workshop (in conjunction with IEEE FG 2011 conference)*, 2011. `http://arrow.dit.ie/dmccon/66/`

- Delaney, S. J. Tarasov, A. Snel, J. Cullen, C. Using Crowdsourcing in the Rating of Emotional Speech Assets. 2011. *International Classification Conference*, St Andrews, Scotland. `http://arrow.dit.ie/scschcomoth/6/`

1 Introduction

1.1 Motivation of Thesis

Human communication through speech conveys verbal and nonverbal information to express intentions, ideas, and emotions. Non-verbal communication through tone of the voice is an important contributing component in emotion expression, which has a significant impact on interpersonal interaction and social influence [18]. As technology becomes increasingly involved in our everyday life, there is a growing interest to achieve human-like communication between people and computers. To achieve more sophisticated interactive systems, much research is being conducting in the domain of emotion recognition in speech. Such research has important implications for the growing development of speech and language research and tech-

nology [19]. Knowledge on how emotion is conveyed in speech is essential for the development and improvement of applications such as intelligible speech recognition systems [20], realistic speech synthesis [21], emotion identification and demonstration in robotics [22], expressive speech animations [23], and interactive applications such as call centres [24].

Despite speech and emotion being intrinsically effortless forms of human communication, they are theoretically and technically complex. It is evident from the ongoing research that both independent and combined domains can be broadly interpreted. Needless to say, it is a fascinating and much discussed topic with many aspects that remain to be investigated. There are a multitude of challenges regarding emotional speech research. For instance, one needs to consider how to conceptualise and operationalise emotion [25, 7, 26]. Indeed, this is potentially the biggest challenge, though it may not appear to be as emotion expression is present in everyday life. However, emotion is not easily defined by simply using linguistic descriptors, rather, it is experienced, expressed, and perceived in a subjective, multimodal, and often ambiguous manner. Many theories exist to explain emotion [27], but no consensual definition or unified approach is currently available. In addition, speech is an acoustically rich and complicated signal. It comprises several communicative functions: linguistic, paralinguistic, and extra-linguistic. These constituent functions are an integral part of speech, each characterised by specific acoustical patterns. However, both general speech acoustics—speaking style and voice characteristics such as gender and age—and affect expressions vary significantly for each individual speaker. This, in and of itself, is a big challenge for speech-related research.

Nevertheless, research has produced many important findings to indicate that emotion can be reliably detected from speech, and uniquely distinguished by acoustical patterns. In more recent times, however, it has become evident that the findings from portrayed speech do not adequately reflect those of natural spontaneous speech—yet, it is not fully understood to what extent such portrayals differ from natural spontaneous speech. In comparison, there is little data on natural emotional expressions available, and at the moment there is a compelling need for further studies using natural vocal expression [28, 29].

1.2 Aims of the Thesis

Many emotion recognition systems emulate the processes of human inferences. To recognise emotion automatically from paralinguistic information, one requires a comprehensive understanding as to the nature of the inference process of emotion from vocal acoustics—irrespective of the intertwining semantic content. Numerous acoustic features indicate emotion in speech, but the extent to which each influences perception of emotion in natural, spontaneous, mood induced speech remains unknown. Arguably, by removing, masking, or manipulating verbal/vocal cues in expressive speech, a listener's perception of expressed emotions should be constricted or misrepresented, therefore, allowing us to quantify its effects. This thesis addresses these issues and attempts to contribute to research in this field by answering the following research questions:

RQ1: What are the practical prerequisites for carrying out large-scale listening tests?

RQ2: Can listeners adequately capture variation of Activation and Evaluation of emotion in naturalistic speech?

RQ3: Can mood induction procedures provide naturalistic speech with sufficiently discernible levels of emotion?

RQ4: Does nonverbal naturalistic speech convey Activity and Evaluation levels that are recognisable to listeners?

RQ5: How do ratings from two perceptually different conditions (verbal and nonverbal speech) compare?

The objective of these research questions is two-fold, though interconnected. First, research questions one and two (RQ1 and RQ2) attempt to determine the suitability of a web-based rating tool using the dimensional descriptive scheme. Research question three (RQ3) investigates the efficacy of Mood Inducing Procedures to provide speech characterised by natural spontaneous emotion. Second, research questions four and five (RQ5 & RQ6) investigate the

effectiveness of *cue modification* techniques to explore the resulting effects on human emotion inferences. That is, these techniques allow us to investigate if listeners can reliably judge expressed emotion in nonverbal aspects of speech.

1.3 Thesis Structure

The work contributing to this thesis is structured as follows:

Chapter 2 - Theoretical Foundations. This chapter deals with the task of conceptualising emotion. It covers several influential theories and research traditions prevalent in the literature. The theories of emotion are grouped into four main perspectives, each making different assumptions about the fundamental nature of emotion. These include the evolutionary, the psychophysiological, the cognitive, and the social constructivism traditions. A general conception of emotion comprising several components is suggested.

Chapter 3 - Emotional Labelling. This chapter is concerned with emotional speech labelling, and reviews the literature relating to assessing and representing emotion. Different methods employed to assess emotion that focus on subjective feelings, physiological changes, behaviours, or cognitions are outlined first. This is followed by a discussion on the issues related to the classification of emotion. The most widely used representations are reviewed, which include discrete and dimensional models. Moreover, a review of recent studies is discussed in relation to descriptive frameworks used for natural spontaneous speech.

Chapter 4 - Emotion and Speech. This chapter outlines the integrated domains of emotion and speech. The topic is described in terms of the Brunswikian lens model, which illustrates the various stages of the communication process between emitting and registering signs of emotion. Accordingly, different research disciplines of the vocal communication process of emotion are reviewed. Moreover, the integration and evaluation of prosodic and semantic cues in emotional speech is discussed. Following this, a review is provided of the issues concerning the different types of speech materials.

Chapter 5 - Acoustic Correlates of Emotion in Speech. This chapter examines certain acoustic parameters that have been correlated with the expression of emotion in speech. The source-filter model is described, following a general discussion on prosody, pitch, time, intensity, voice quality and spectral-related features.

Chapter 6 - Development of an Online Rating Tool. This chapter details the development of a web-based rating tool to gather ratings from large-scale listening tests, with the design placing particular emphasis on participant accessibility. This chapter also details amendments made to the tool in order to be suitable for experimental work.

Chapter 7 - Case Study: Labelling Mood Induced Speech. This chapter details the first perception test carried out to obtain ratings for labelling emotion in a naturalistic speech dataset. The ratings suggest that mood induction procedures can provide sufficient levels of emotion in naturalistic speech. The ratings provided labels to complete the build of a naturalistic emotional speech corpus.

Chapter 8 - Judging Emotion from Low-pass Filtered Naturalistic Speech. This chapter covers the experiments that investigate listener inferences based on intact and non-verbal cues in two conceptually different speech stimuli. This chapter examines a smaller derived speech dataset based on high agreement levels obtained from ratings in the case study of Chapter 7. A survey was carried out on different filtered stimuli to establish a suitable condition to administer. The results show that low-pass filtering has a relatively small impact on the perception of Activation and Evaluation.

Chapter 9 - Conclusion. This chapter concludes the research and summarises the contributions made in this thesis. The rationale for each aspect of the research are explained in respect to the research questions asked throughout. Finally, future work is considered to compliment the research undertaken in this thesis.

2

Theoretical Foundations

An essential requirement in undertaking the research in this thesis is the task of labelling collected speech data for affective states. There are many challenges with regard to labelling of emotional content, both technical and theoretical. On the face of it, emotion may seem easy to define because everyone encounters it in everyday life. However, the nature of it is a topic of on-going debate of which there is a large body of literature spanning many disciplines (e.g. philosophy, psychology, sociology, cognitive science) and a confirmed definition is difficult to obtain. In fact, contemporary researchers [30, 25, 31, 32] explain that there is no consensually established methodology for describing emotion or labelling data that signify it. One must carefully plan an approach to labelling emotion in data, and to do so a comprehensive understanding of emotion—or more specifically an awareness of the available theories—is necessary. A researcher investigating the nature of emotion is confronted with a variety of definitions and

theories, and many viewpoints can be considered. It is not essential to be familiar with all theories and definitions relating to emotion. Nevertheless, theories steer the orientation of a specific piece of research as they allow us to make informed decisions about practical implications. This chapter briefly covers some of the different theories available from which labelling methods have emerged.

Theories of emotion can be categorised in terms of the underlying nature within which they are claimed to be best understood. The four perspectives are: the *Darwinian, Jamesian, cognitive* and *social constructivist* [27]. Because each perspective makes different assumptions about the fundamental nature of emotion, there are various approaches to defining and examining it. This can make it difficult for findings to be integrated. Although the perspectives do at times reconcile, at other times, they are contradictory. It is, therefore, important to be aware of the various perspectives in which a study may belong.

2.1 The Darwinian Perspective

The Darwinian perspective is concerned with evolutionary psychology, which claims that the best way to understand emotion is to view it in terms of psychological responses for survival. In 1872, Charles Darwin wrote in his book, "The Expression of Emotion in Man and Animal" [33], that emotion expressions are vestigial reaction patterns shaped through evolution to assist a basic survival function, i.e. *adaptations* shaped by natural selection [34]. This theory not only suggests that all humans share certain basic or fundamental aspects of emotion, but also that certain aspects are shared with other related species [35, 36, 34, 37]. The implication of this theory has prompted animal research and has gained substantial experimental evidence for the link between human and animal neurophysiology of emotion. Panksepp mentions that to the best of their knowledge, "basic affective feelings supervene on homologous brain systems shared by all mammals" [38]. The notion that certain emotions are innate, as psychological and biological, amongst all humans, suggests that the same emotions should be expressed, to some degree, instinctively across different cultures. To investigate this, a cross-cultural study

by Ekman [39] provided some support for this theory. His work focused on the universality of facial expressions for a small set of *basic* emotions [40]. The most widely accepted of these basic emotions are happiness, sadness, anger, fear, surprise and disgust—often referred to as the "Big Six" [41] (discussed further in Chapter 3.3.1).

If there are a certain number of basic (or primary) emotions evident in facial expressions, one would expect to see similar evolutionary indications in other modalities of emotion, such as vocal expressions. As it happens, Darwin [33] provided the first thorough discussion of vocal emotion expression [42] and suggested the correspondence of vocal expression with the speaker's emotional state. A study by Scherer et al. [43] suggested that there are similar inferences of emotion from vocal cues across nine different countries—from content-free speech. However, accurate detection decreased as the participants' language became more dissimilar from the presented language, i.e. German. As acknowledged in their study, intercultural studies may be susceptible to criticism due to contemporary mass media, such as Hollywood films and the associated familiarity of particular types of expressions. They argued, however, that vocal expressions by German actors are unlikely to be influenced by Hollywood films because most foreign films are usually dubbed in German. To deal with common exposure as a potential artefact, Bryant et al. [35] conducted a study examining the perception of vocal emotional expression in a South American indigenous population with relatively little exposure to common sources of emotion stimuli. The study demonstrated that two relatively different cultures could reliably identify basic affective categories—happiness, anger, fear, and sadness—from emotional vocalisations, and thus attribute universal recognition to global acoustic properties such as speech prosody.

2.2 The Jamesian Perspective

Philosophers have questioned the nature of emotion as far back as Socrates and even "pre-Socratics". According to Solomon [44], Aristotle also discussed certain emotions at length, notably the emotion of anger. His examination of anger and his philosophical reflection on

reason centred on the cognitive aspect of emotion, yet acknowledged some sort of accompanying bodily distress. Similarly, René Descartes [45] acknowledged bodily attributes, implying that emotion requires the interaction of the body and mind. He termed the bodily sensations associated with emotion as "animal spirits", which refer to agitation when the mind and body meet at a small gland at the base of the brain (the pineal gland). Other subsequent philosophers, such as David Hume [46], have also interpreted emotions as a certain kind of physical sensation—which he called an "impression".

2.2.1 James-Lange Theory

The connection between emotions and physiological changes in the Darwinian tradition is central to subsequent theories of emotion in the Jamesian perspective. Darwin's perspective, however, was more concerned with emotional *expression*, whereas William James was concerned with emotional *experience*. In his 1884 article "What is an emotion?" [47], James inspired theory and research in the physiological realm of emotion. He argued that emotional experience arises from the perception of bodily changes. Around the same time, the Danish psychologist Carl Lange proposed a similar theory from which the *James-Lange* theory arose. The James-Lange theory suggests that emotional experiences arise from the perception of bodily changes when evaluating an external stimulus. It proposes that the physiological reactions are primary and that our interpretation of those reactions, which elicits an emotion, is secondary. Their theory produced later work on the *facial feedback hypotheses* [48, 49, 50, 51]. The facial feedback hypothesis shows that a certain facial expression will cause a change in a person's emotional state. Strack et al. [51] reported that participants who had unknowingly been induced to a smile by holding a pen in their mouth found comic strips funnier than participants who had not. Even in the absence of visible facial expressions, using facial EMG (electromyography) it has been demonstrated that there is a direct link between facial expressions—motor schemata of facial muscles—and emotions [52, 53]. For example, recordings have shown that participants respond faster to pleasant and unpleasant words read on a screen when smiling and frowning, respectively [54]. A study by Barrett et al. [55] established that interoceptive cues (as measured by heartbeat detection sensitivity) are linked with reports of experienced emotion, but that these

cues are more important for some individuals than for others. The study suggests that participants, who more accurately perceive their own heartbeat in a detection task, emphasise feelings of activation and deactivation in self-reports of emotion to a greater extent.

2.2.2 Cannon-Bard Theory

As mentioned, the James-Lange theory states that physical arousal precedes emotion. This notion was radically challenged by Walter Cannon [56] and later with Philip Bard. Together they developed the Cannon-Bard Theory. Unlike the theory proposed by James and Lang, Cannon and Bard contend that physiological changes are a byproduct of emotions, and that these changes occur simultaneously with emotional changes in response to a stimulus. Other commentators support the notion that there is no significant influence of autonomic events[1] on emotion. For example, it has been shown that patients with spinal-chord injuries or total autonomic failure exhibit little, if any, impairment in the experience of emotion [57, 58, 59]. In brief, a major role of this theory is associated with thalamic processes in the brain, which inspired early functional neuroanatomical models of emotion [60, 61].

Debate on the theories of James-Lange and Cannon-Bard centres on *peripheralism* versus *centralism*. Peripheralists (James-Lange) maintain that the various peripheral bodily changes, such as heart rate, contribute to the emotion, whereas Centralists (Cannon-Bard) maintain that peripheral activity is irrelevant, and that emotion is primarily a function of the brain. Both theories are radically opposed with regard to the temporal sequence of an emotional event, of which a detailed account is beyond the scope of this research. Fundamentally, however, both theoretical frameworks are concerned with the physiological processes (somatovisceral activity) resulting from an experienced emotion.

[1] The autonomic nervous system is the part of the peripheral nervous system responsible for controlling bodily functions that function largely below consciousness, such as breathing, the heartbeat, and digestive processes.

2.3 The Cognitive Perspective

According to Dalgleish and Power [62], the cognitive perspective of emotion is considered by many researchers to be the fundamental component of emotion. It assumes that thought and emotions are inseparable and it is evident that it is deeply incorporated into the three other perspectives [27].

2.3.1 Schachter–Singer Theory

Schachter and Singer [63] revised both the James-Lang and Cannon-Bard theories, and developed their own, now also known as the *two-factor theory*. Although partly concurring with the ideas of both, they contended that physiological arousal can instigate emotions but that physiological changes are not specific to a particular emotion. They proposed that the emotion is determined by the interaction of physiological arousal and cognition of the event triggering the physiological arousal. In other words, the theory proposes that an affective response is contextually driven, and that cognition determines whether the physical arousal indicates a state such as 'anger', 'joy', etc.

Damasio's [64] somatic marker hypothesis and Barret's [65, 66] somatic marker and conceptual act model are in line with the conceptualisations of James and Schachter and Singer. Damasio, a neurologist, introduced the idea of *somatic markers* (physiological reactions). This hypothesis proposes that emotions arise from bodily experiences which influence behaviour, and especially that of decision-making. Somatic markers provide a means for evaluating which current events had previous emotion-related consequences [67]. Barret recently adapted Schachter's theory of emotion with the *conceptual act model*, a social neuroscience model of emotion. It hypothesises that discrete emotions emerge from a conceptual analysis based on the notion of *core affect*[2] put forward by Russell [68]. The model challenges the position of *basic* emotions being biologically hardwired into the brain.

[2]Core affect is characterised as the momentary changes in an organism's neurophysiological state that represent an immediate relationship to the flow of changing events [66].

2.3.2 Appraisal Theory

In cognitive emotion theories, Magda Arnold [69] is accredited with the *appraisal theory* and its modern approach, although the origins of this perspective stem further back. She proposed that people evaluate events in their environment, judged as good or bad, resulting in an appraisal of a stimulus and hence perceived as an emotion [69, 27]. It argues that every emotion is associated with an appraisal influenced by the individual's learning history, personality and physiological state. The appraisal hypothesis is closely related to the idea that emotions are *action tendencies* [70]. The process of appraisal informs the organism of the particular features of a given environment and brings about a state of readiness to act on those features.

2.3.3 Lazarus Theory

In line with Magda Arnold's appraisal theory, the concept of cognitive appraisal is central to Lazarus' theory in which he deals with emotions and coping. The theory states that cognition occurs before any emotion or physiological arousal. Lazarus specified two major types of appraisal methods: (a) primary appraisal, which is the evaluation of the significance or meaning of an event, and (b) secondary appraisal, which is directed at the evaluation of the ability to cope with the consequences of that event.

2.3.4 Cognitive Emotion Models

The component process model of emotion, developed by Scherer [71], is based on the sequential check theory of emotion differentiation (a set of criteria that are predicted to underlie the assessment of the significance of a stimulus event for an organism). It implies that bodily expression is an outcome of cognitive appraisal but disregards the primacy focus of cognition, considering it as a constitutive synchronous element. The model attempts to explain how a sequence of specified stimulus evaluation (appraisal) checks bring about differentiation of emotional states [72]. This framework considers five crucial components that are synchronised for a short period of time, where the components of an emotion episode represent the respective

states of all major organismic subsystems. The five components are: cognitive (appraisal), neurophysiological (bodily symptoms), motivational (action tendencies), motor expression (facial and vocal expression), and subjective feeling (emotional experience) [7]. Based on his theory, Scherer [73, 7] makes detailed physiological predictions of vocal changes associated with specific emotional states determined by particular profiles of appraisal outcomes.

Ortony et al. [74, 6] developed a cognitive emotion model that is computationally tractable, called the *OCC model* (Ortony, Clore and Collins). In the OCC model, emotions are valenced reactions to three types of stimuli: consequences of *events*, actions of *agents*, and aspects of *objects*. The goal of the model is to predict and explain under which circumstances certain emotions are likely to occur. This model is explained in more detail in the next chapter (section 3.3.2).

Central to the cognitive perspective evaluation of stimuli is how one "appraises" events in one's environment. Cognitive theorists, in contrary to the perspective of James, consider cognitive processing of stimulus information as primary and the physiological components associated with emotion as secondary.

2.4 The Social Constructivist Perspective

In contrast to the Darwinian view of the evolutionary development of emotions is the idea that emotions are socially and culturally constructed. From the social constructivist perspective, all emotions and human behaviour are learned by individuals through experiencing social and cultural rules [75]. Typically, social constructivists suggest that emotions are best understood as the products of learned social rules, necessary for successful social interaction, and that the biological foundations of them are of secondary importance. Averill [76], for example, argues that emotions cannot be explained solely from an evolutionary or a physiological perspective, and suggests they can only be fully understood in terms of the social functions they serve [27]. He describes emotions as socially determined cognitive appraisals that result in behavioural

"scripts". How anger is expressed, and what causes anger to emerge, for instance, differs from culture to culture. While Darwinians typically support their position with the apparent "universality" of emotions across different cultures, social constructivists often support their position by reporting the discrepancies of words related to emotion found in languages of different cultures. Furthermore, social constructivists have shown that there are no obvious bodily correlates with subtle or advanced emotions, such as love and guilt [77]. It is also argued that there can be a noticeable change in expressed emotions concurrent with the change of essential social functions [76, 78]. In other words, some emotions may culturally disappear with new ones emerging [79].

While both strands of research appear to have strong evidence to support their respective positions [77], the views in each tradition are not mutually exclusive. Even well known Darwinians, such as Ekman [80], acknowledge the role that culture plays in regulating emotional displays, and most social constructivists acknowledge that emotions are, to some degree, innate. For example, some social constructivists agree that strong positive and negative affects may be an inborn structure of the *human* brain [65, 68]. In contrast to Darwinians, however, they rarely consider this to be the case for *mammalian* brains [81]—this contention awaits neurobiological support [82]. A study by Matsumoto et al. [83] amalgamated both perspectives, with the notion that an individual's emotional displays can be both universal and culturally variable. Their study examined how time-dependent emotional displays of Olympic athletes vary between different cultures. They found that an athlete's initial emotional expression tended to be universal, while over time subsequent expressions became culturally regulated. The variance between athletes was determined by whether the athletes were from an individualistic culture, in which they appeared more expressive, or from a collectivistic culture, in which the expressions appeared more concealed.

2.5 Emotions as multifaceted

It can be seen that within each perspective the study of emotion is approached according to the different basic assumptions made about its nature. Although they may differ in their methods, certain principles do coincide. The four perspectives discussed focus on different aspects of emotions but there is evidence that they have begun to converge [27, 41], most apparent to researchers of the Darwinian and Jamesian perspectives. The Darwinian perspective, dealing with evolutionary theory, and concerned with 'fundamental' or 'basic' emotions, has been adopted by both Jamesian and Cognitive researchers, as can be seen in much of Ekman's and Lazarus' work. The first three perspectives (Darwinian, Jamesian and Cognitive) seem to integrate into a more cohesive view of emotion encapsulating the different research paradigms. The social constructivist perspective partly conflicts with the Darwinian perspective, proposing emotions to be socially learned [76]. Against this stance, some evidence suggests that non-Western congenitally blind individuals, including infants, display emotional facial expressions similar to Western individuals, indicating the possibility that emotion is not merely obtained via social learning [84, 85, 86]. Although the social constructivist perspective may not comprehensively reconcile with the Darwinian and Jamesian perspectives, it does incorporate certain aspects provided by them [27]. Each research paradigm, evolutionary, neurological, biological, psychological or sociological, uses different methodologies that consider the different perspectives. Contingent on the particular perspective, certain descriptive frameworks emerge to meet the requirements of a particular aspect that a theory emphasises [87].

Theories have been proposed over a wide range of disciplines, each with a focus on a particular manifestation or component of emotion. At present, it is commonly agreed that emotion is best conceptualised as a multifaceted phenomenon that consists of several components/elements [88, 7, 16, 89, 90, 91, 70]. The most commonly accepted of these elements are:

- subjective experience (feelings);
- physiology (body activation);

- expression (facial and vocal expressions).

Mauss and Robinson [88] provide a review for measuring emotion that focuses on the above "emotional responses" which arise from an appraisal of a situation, rather than focusing on cognitive antecedents and their respective correlates of emotion. They refer to "behaviour" as opposed to "expressions", which takes into account the link between emotional states and the tendency to act on them [70]. Scherer's [7] suggestion for a comprehensive measuring system, however, includes the notion of evaluating objects and events (appraisal) and the disposition to act (behaviour) as separate assessable components—albeit somewhat less consensual. These are referred to as:

- the cognitive component (appraisal);
- the motivational component (action tendencies).

In order to explicitly represent emotion in research, one needs to consider the theories that representational issues are guided by. These issues are discussed in the next chapter.

3

Emotional Labelling

3.1 Introduction

To conceptualise emotion, the last chapter gave a general overview of the different types of theories that are well-known in psychology, emphasising core aspects that represent the nature of each perspective. These theoretical foundations lead to the practical considerations for assessing and representing emotion, two essential matters involved with emotional data labelling. The aim of this chapter is to give the readers an overview of the considerations that are made when labelling speech data for emotion. As mentioned in the last chapter, emotion is typically conceptualised as being multifaceted that involves synchronised changes in several constituent elements [92, 7] (see section 2.5). Scherer [7] suggests that in order to provide a comprehen-

sive representation of emotion, all components should be measured and assessed to evaluate their convergence. Each component, however, is distinct in nature, and there is currently no technique that amalgamates them all. Nevertheless, one should keep Scherer's point in mind—the whole is greater than the sum of its parts. Therefore, in order to minimise methodological inconsistencies between our approach and other measurement types, first, it is necessary to familiarise ourselves with the different assessment techniques, and second, it is crucial to consider carefully which method to use to explicitly represent emotion that is capable of distinguishing between distinct emotional states.

3.2 Emotion Assessment

In any scientific investigation some kind of measurement is required. In relation to emotion, finding appropriate measurement methods is complicated, and mainly because emotion is an abstract notion difficult to clearly define. Research involving the measurement of emotion has a long history and it has proven to be a considerably complex task. Typically, measurement methods are established according to one or more of the constituent elements of emotion—cognitions, subjective feelings, physiological changes, and behaviours. These measurements are as follows:

- **Self-reports:** are the participants' description of their subjective feelings, cognitions, and, sometimes, behaviours typically obtained from open-ended questioning or psychometric instruments.
- **Physiological measurements:** are recordings of physical changes that occur during an emotional episode. These include responses in heart rate, blood pressure, skin conductance, pupillary dilation, respiration, and brain waves.
- **Behavioural observations:** are the emotion inferences of observable actions in response to an emotion, such as facial, vocal, or whole body expressions. Behaviour/expressions are generally assessed by objective observers in judgement studies.

This section first examines each assessment type and discusses to what extent these measurements may cohere. In providing labels, it is also essential to report on the description of the labels provided, depicted as cause-or effect-type orientation. Moreover, we will briefly discuss the rationale behind the selection of judges who will effectively assess speech for labelling.

3.2.1 Self-reports

The use of self-reports is a popular method in emotion research, and it serves as the basis for a lot of the available evidence about emotional experience. It is favoured for several reasons. This method is relatively straightforward, cheap, and can be easily administered to large sample groups—delivering it via the web, for instance. Moreover, self-reports allow a researcher to evaluate whether the objectives of eliciting emotion can be met by systematic inducing procedures, or by being exposed to a specific set of stimuli. Typically, multiple assessments are used to gain information. Self-reports are used to link the experiential component with physiological, behavioural, or both components. In speech research, for instance, labels obtained from self-reports (cause-type descriptions) can be linked to vocal indicators of emotion to determine the acoustic correlations.

Self-reports are the only means to gain information about the experiential component of emotion [7]. However, they are considered subjective in their own right and are likely to be unreliable [16] due to various psychological factors. For example, Robinson and Clore [93] argue that self-reports of retrospective emotional experiences—which partially reflect recalling contextual details of an event—are less likely to be valid than self-reports about current emotional experiences. This suggests that self-reports should be administered during, or at least shortly after the emotion is being experienced. This can, however, be problematic if it interferes with the emotion induction procedure, making the effect of the inducing procedure, or evoking stimulus potentially uncertain. Moreover, it has been shown that self-concepts are influenced by gender stereotypes, which as a result differentiate individual self-knowledge [94]. Self-reports rely on the participant's own assessment of the emotional experience, so they are

reliant on a participant's honesty, introspective ability, and conceptual understanding of emotion descriptions—dimensional concepts of emotion, for example, may not be understood by a lay person.

3.2.2 Physiological Measurements

As mentioned, physiological measurements are concerned with the body changes that occur when a person experiences an emotional state. While all traditions acknowledge bodily activation, those most concerned with the physiological aspect of emotion are the Darwinian and the Jamesian perspectives. Bodily changes can be assessed using psychophysiological instrumentation and methods that are concerned with the link between the psychological meaning—higher cognitive processes—and physiological responses. Measurable changes occur in both the central nervous system (CNS) and the peripheral nervous system (PNS). The CNS comprises the brain and spinal cord, which send and receive signals from the peripheral nervous system. The origins of examining brain activity associated with emotion can be affiliated with the development of Cannon and Bard's theory (see section 2.2.2). Researchers in this domain are particularly concerned with how emotions are embodied in the brain, rather than the physiological responses in the peripheral system [95, 67]. Techniques for measuring brain activity—which can be administered in parallel with other assessments of emotional speech—include electroencephalography (EEG) [96, 97, 98], magnetoencephalography (MEG) [99, 100], and neuroimaging techniques, such as positron emission tomography (PET) [101, 102], and functional magnetic resonance imaging (fMRI) [103, 104, 105]. With EEG measurements, electrodes are attached to the participant's scalp to measure the electrical activity of cells. The biggest advantage of EEG is its temporal resolution, but it is limited in detecting the precise location of activity (spatial resolution), thus, it is used to measure activation in relatively large areas. To achieve both high temporal resolution and spatial resolution, EEG can be combined with MEG data [106]—MEG records magnetic activity rather than electrical activity. Neuroimaging studies, using fMRI—measuring changes in oxygen intake—and PET—measuring a radioactive isotope injected into the bloodstream—are more successful in locating activity in specific brain regions. Unlike EEG, fMRI and PET involve using large expensive machinery

and typically requires induction methods where participants are asked to recall a past event. In many cases, it may be difficult to determine if particular brain activity is associated with a recalled memory of an emotional experience, or simply the cognitive task of recalling the event itself [107].

Located outside of the brain and spinal cord, the peripheral nervous system (PNS) is subdivided into the somatic nervous system (SNS) and the autonomic nervous system (ANS). The functions of the SNS include the voluntary control of the skeletal muscles, which comprise those of the face, and the carrying of sensory information (e.g. touch, temperature, and pain). Studies that link motor expressions (e.g. facial and vocal) with emotions are central to the work of the Darwinian tradition. Because expression converges with behaviour, we discuss it in a separate section (section 3.2.3).

The ANS primarily regulates the internal environment of the body, which includes involuntary activity, such as skin conductance, heart rate, pupil dilation, and respiration. Much of the current research that employ ANS measurements is inspired by James' theory (see section 2.2). Within the ANS there are two subsystems: the sympathetic nervous system and the parasympathetic nervous system. The sympathetic nervous system prepares the body for action in threatening situations (i.e. fight or flight), while the parasympathetic nervous system acts to restore bodily functions, such as slowing the heart rate down. The characteristics of the anatomical and functional changes within the parasympathetic and the sympathetic branches have been important in the study of emotion. Measurements of physiological responses of emotion include heart rate (HR) [108], cardiac output (CO) [109], electrocardiography (ECG) [110], blood pressure (BP) [111], facial electromyography (fEMG) [112], pupillary dilation [113], respiration rate [114], and galvanic skin response (GSR) [115]. Psychophysiologists have built up a wealth of empirical evidence to support the relationship between physiological measurements and emotion. Several extensive reviews on emotion and both nervous systems have been provided [116, 117].

One of the most obvious advantages of using physiological measures is their inherent objectivity, and the ability to provide continuous measurements. However, there are some limitations. For example, many of the instruments used are expensive, and some are obtrusive. Moreover, physiological signals are sensitive to other bodily functions associated with multiple mental or physical activities that are not necessarily a function of emotional response, such as functions related to digestion, homeostasis, exertion, etc. [88]. Even with careful control of intruding variables, physiological changes are highly context-dependent, and correlations with emotional states have been shown to be inconsistent at the best of times [118].

There have been an ample amount of studies conducted relating to emotion in speech that simultaneously integrate information from physiological signals of different modalities [119, 110, 111, 113, 115, 120, 121, 122]. Generally, physiological response measures are linked to the subjective experience—assessed by self-reports [110, 119]—or to the perceived expressions—assessed by judgement tasks [123, 113]. The study by Johnstone and Scherer [121] obtained both physiological and acoustical data and correlated this with self-report data, obtained from participants who were asked to report how they were feeling at the precise time that the induction procedure (game play) took place—addressing the notion that recalling an emotional experience may not be as valid as reporting on current experience. On the other hand, some studies make no account of self-report or observable data. Instead, they may make the assumption that the induction procedure successfully elicits the intended emotion [122], or, explicitly, associate physiological responses with a specific stimulation [112, 124].

3.2.3 Behavioural Observations

Needless to say, nonverbal behaviour (expressions) is a key component of emotional responses. As previously mentioned, Darwin proposed that expressive behaviour was the outcome of evolutionary development that exists to serve communicative functions. In line with behaviourism, Frijda [70] considers emotions as "action tendencies", tendencies to engage in behaviour as a result of an evaluated stimulus. He argues that emotions emerge because of an individual's

appraisal of an event and its relative value for well-being, which, in turn, prepares "action readiness" manifested as expressions. Studies of behaviour are concerned with body, facial, and vocal cues, and can operate independently of a spoken language system, as studies concerning newborns [125], indigenous peoples [35], and animals [126] have demonstrated.

Body Cues

To investigate how emotion is communicated through the body, one can either analyse dynamic movement (e.g. motion capture systems) [127] or static postures (e.g. photographs or figures) [128]. Body movement, or "action" behaviours, usually have relatively distinct beginning and end points, which involve functions such as nodding, shrugging, and gesturing [129]. Action tendencies can be conceptualised as responsive actions as a result of an appraisal process. In effect, approach and avoidance behaviour can be used to gauge the observed hedonic-evaluation of a particular stimulus. Studies show that participants who evaluate stimuli as pleasant, as opposed to unpleasant, produce approach tendencies, and negative evaluations produce tendencies to avoid [130]—although the actions depend on context and on the participant's desired goal [131]. On the one hand, it has been shown that recognition rates from body postures, for specific emotions such as anger, sadness, and happiness are comparable to that of the voice and face—yet accuracy rates for disgust, for example, are comparatively low and it is understood to be primarily communicated through the face [128]. On the other hand, it has been suggested elsewhere [129] that body movements alone cannot portray specific universal emotions. Instead, body movements are believed to contribute to, or emphasise expressions in line with facial and vocal cues, and allow information to be communicated about aspects such as attitude and status. A study by Nelson and Russell [132] showed that bodily expression is essential in multi-cue emotion communication. They investigated children's understanding of nonverbal cues of pride. They showed that 4- and 5-year old children did not recognise pride from facial or bodily cues independently, nor from integrated face and body cues. Even 6- and 7- year old children could only infer pride from combined body and face cues, but not from face or body cues separately. Influenced by the facial feedback hypothesis (see Chapter 2), body postures are also believed to have an impact on experienced emotion. When considering pride gained

through the achievement of a task, it has been shown that sitting in upright postures led to greater feelings of pride compared to sitting in slumped postures [133].

Facial Cues

Presenting facial stimuli is an important method in emotion perception studies. Similar to the study of body behaviour, stimuli can be presented as either dynamic [134] or static [90]. So far, it appears that most studies have been limited to static stimuli, e.g. the use of photographs. As mentioned previously, much work on facial expressions [40, 135, 86] can be affiliated with the Darwinian perspective of evolutionary theory and the concept of having a few 'basic' emotions. Ekman's findings on prototypical facial expressions suggested at least six 'basic' emotions. Based on facial anatomy, his work [85] resulted in the development of a systematic observation tool called the Facial Action Coding System (FACS, for more detailed information see section 3.3.3). While fEMG and FACS measure activity of facial muscles, EMG measures the electrical potentials from muscle contractions, which may not be visibly discriminative by judges. Because EMG involves placing electrodes on a subject's face, it is somewhat invasive. This may cause confounding effects, such as increased self-conscious behaviour. EMG has a higher level of granularity of contraction measures, but both are said to be highly correlated [136]. Although researchers employing FACS measurements require in-depth training, many use this method successfully to link induced emotional states to distinct facial expressions (see [137] for a comprehensive review).

A direct link with facial expressions seems problematic and far from straight forward. A smile, for example, can be associated with both happiness and nervousness, or with either failure or success [138]. A review provided by Russell et al. [139] suggests that smiles are often limited to social circumstances—rather than directly related to a happy event. Although links have been shown to exist between emotions and certain facial behaviours [140], according to Mauss and Robinson's review [88] the conclusions made about the relationships between discrete emotions and distinct facial response patterns are questionable (see also [139]). They argue that facial expressions are more reliably linked to the hedonic valence of a person's state.

Vocal Cues

Finally, expressive behaviour is also known to exist in vocal communication. In this domain, researchers consider an expression as the qualities of speech that are communicated through non-verbal content, also known as *paralinguistic* content. It is difficult to determine the boundaries to which listeners infer emotion from either paralinguistic or semantic (lexical) content. Decoding studies have shown that listeners are able to infer—actual or simulated—emotion from paralinguistic cues irrespective of lexical content [35, 141]. Many studies use actors to portray a set number of different emotions to produce speech utterances composed of standardised or nonsensical content [142]. This way the speech does not contain any linguistic information that could indicate the underlying emotion of the speaker. Alternatively, for non-acted emotions one can remove semantic content by cue masking. Cue masking is used to distort speech, by removing or altering certain acoustic cues, to make speech unintelligible. Techniques that are used include low-pass filtering, randomised splicing, playing backwards, pitch inversion, and tone-silence coding [143, 144, 145, 146]. It has been shown that specific acoustic vocal cues are associated with discrete affective states [73, 147], yet can also be measured along a small set of continuous dimensions [13, 16, 148]. Because this thesis is concerned with emotion in speech, we will dedicate a separate chapter on this topic in order to cover it in greater detail (see Chapter 4).

3.2.4 Coherence between Emotion Components

In componential emotion theory, it is postulated that coherence across multiple components (cognitive, behavioural, experiential, and physiological) occurs in response to an emotional episode [149]. Scherer [7] explains that on the basis that all subsystems underlying emotion components—such as the central nervous system accounting for subjective feeling and appraisal, and the somatic nervous system accounting for motor expression—are typically independent, hypothetically they become temporarily interrelated and synchronised during an emotional episode. However, there are contradictory findings on the coherence among the components [91], and further investigations are required. A study by Bonanno and Keltner [150],

for example, compared spontaneous facial expressions and emotion-related appraisals with experience of emotion. They found moderate associations with facial expressions and appraisal profiles of anger and sadness, yet these results were greater in degree compared with smiling and laughter. In fact, their data showed that laughter occurred more often during appraisals of anger. Similarly, Mauss et al. [92] found that three components, experience, facial behaviour, and peripheral physiology, are indeed associated but that coherence varies in degree. The results indicated that the experiential and behavioural responses are highly associated, while the physiological responses are only modestly associated with experiential and behavioural responses. One would assume, however, that behavioural and physiological measurements are mostly inter-dependent, and less discordant compared to reports of subjective experience and physiology [117]. Moreover, a study by Gentsch et al. [149] investigated the synchronisation of specific appraisal processes, as predicted by the component process model [151], simultaneously with facial muscle activity and brain activity. The results of the appraisal-driven response changes supported the predictions made by component process model. However, they recognise that the generalisability of the results are limited and that investigating coherence is eminently challenging. They acknowledge, for example, that temporal dynamics, context, and individual differences (see also [88]) bring several complexities.

It seems that across different contexts, components are only loosely coupled [139], as is evident in light of the recent findings. The review provided by Mauss and Robinson [88] demonstrates that different measures of emotion only correlate low to moderate. They suggest that "one measure of emotion is likely associated with variance unique to [that component]". The example that they provide suggests that facial EMG is more responsive to valence, while skin conductance is more responsive to arousal. Similarly, with voice communication it has been shown that valence is more difficult to detect compared to arousal [152]. In summary, Mauss and Robinson suggest that there is no 'gold standard' for measuring emotional responses (see also [7]). One should treat emotion as a multifaceted phenomenon, and try to interpret each facet and its constituent role.

3.2.5 Cause-type and Effect-type Descriptions

In most cases, the investigation of emotion involves linking one component of an emotional reaction with another. The analysis of speech and emotion, for example, requires correlating acoustical measurements from the speech waveforms (physiological changes of voice) with either the measurements obtained from self-reports (experiential component), the measurements obtained from listening tests (behavioural component), or, in some cases, with other physiological responses such as heart rate. In other words, acoustical data is correlated with a *label* that represents either the experiential or behavioural component. In relation to labelling, a valuable distinction between the two emotion label descriptions has been made by Cowie and Cornelius [81] as *cause-type* and *effect-type*[1]. The distinction specifies whether the given label is focused around the realisation of the speaker or the listener. Both types are valuable descriptions, but are required for different applications. Firstly, when describing emotion, a researcher might be interested in the internal emotional state (subjective feeling) of the speaker, which subsequently leads to the speaker to produce emotion-specific vocal characteristics. In this regard, the labelling task sets out to capture information about the speaker's state at the time of speaking (encoding). Research with this goal in mind is referred to be as *cause-type* orientation. An example of cause-type labelling is described in the study by Fernandez and Picard [153]. In this study, the chosen labels (categories) represented stress levels, which were operationalised by four different experimental conditions corresponding to cognitive load. In any case, they considered this approach suitable for the application in mind, i.e. detecting stress under driving conditions.

The second type of orientation is referred to as *effect-type*. It describes the listener's interpretation of a speaker's state based on the characteristics of speech, irrespective of the speaker's truly felt state—or at least the speaker's truly felt state while speaking is of secondary importance [154]. Generally speaking, discrete categories and dimensions are used, which are suitable for the understanding of a layman. Practically, this is the more appropriate type for describing

[1]This related distinction is also recognised by Scherer [12] in the context of the Brunswikian model (see Chapter 4.1.1 as 'encoding studies' and 'decoding studies', and by Schr"oder [87] as 'speaker-centred' and 'listener-centred', respectively.

observable behaviour. Because a speaker often disguises or misrepresents what he or she is truly feeling, it is difficult to provide reliable cause-type descriptions. Inter-rater measurements do not offer any indication of validity for cause-type descriptions [155], but they do for effect-type labels, which makes validation relatively straightforward [156]. An example of labelling specifically identified as effect-oriented can be found in the UU Database [157]. Based on the psychological background, they annotated speech data for expression with six dimensions related to personal emotional state, interpersonal relationship, and attitude. To illustrate rater consistency, they reported the standard deviation values, and for rater agreement they reported the Kendall's *W* coefficients.

Labels that are provided by the speaker's themselves through self-reports are a justifiable description for cause-type, but as mentioned earlier, this method is considered by many as too subjective. When focusing on verifiable properties of a speaker's state through physiological measurements, however, descriptions would typically reflect cause-type orientation [154]. In summary, the important distinction is that what is felt by a speaker does not necessarily correspond to what is being expressed. Cowie mentions that 'cause' and 'effect' interpretations diverge in cases such as deception and acting.

3.2.6 Selection of Judges

To provide labels for emotional speech, the assessments performed in this thesis are concerned with the realisation of the listener, rather than the speaker. The labels are, in other words, of effect-type description, derived from the evaluations of one or more 'judges' of emotional expressions. There are two types of judges that can annotate: "expert" and "naïve". In most cases, expert judges are assigned, which usually comprises a small group that annotates a large number of speech samples. Although it is often not indicated what expertise the judges actually have, in many cases experts are researchers who are part of the wider field of emotional research. In the expert-based approach, issues with the semantics of emotional categories can be solved by agreeing on the meaning of words prior to the task [26], increasing the likelihood

of labelling consistency. However, although expert ratings may yield reliability, they may not yield validity because labelling would be representative of theory. Moreover, expert rating may even embody theoretical bias [158]. Because the nature of the labels provided are based on the nature of the population of the judges, one may, on the other hand, be interested in ratings that are representative of a broader sample population, including judges who are not necessarily familiar with emotion theory, i.e. "naïve" judges. Emotion is, after all, an important aspect of communication between *all* humans. Instead of using a small number of expert raters, large-scale listening groups have also been proposed [159], or the more recent phenomenon known as crowdsourcing [160], to provide labels for emotional speech. Crowdsourcing is the process of outsourcing tasks to a large group of undefined non-expert individuals [161]. In the context of labelling speech corpora, each speech clip is presented to several raters and rated separately by each individual. The final label for a speech clip is some combination of these ratings. Crowdsourcing has recently been used for labelling corpora in numerous domains, such as machine translation [162], computer vision [163, 164], and sentiment analysis [165, 166]. Crowdsourcing is a fast and effective way of accumulating ratings [166], yet can provide the same quality labels compared to those provided by expert judges [167] if a sufficient number of raters is acquired.

3.3 Representing Emotions

Regardless of which component of emotion is studied, it is essential to find an appropriate way to describe it. For this thesis, in essence, to investigate links between speech and emotion-related states, we need suitable methods to represent the emotions that are portrayed in speech. Due to the complexity and uncertainty of the emotion phenomena, describing and labelling emotion is not a straightforward task. According to the literature, currently no generally accepted methodology for labelling exists [31, 10, 16]. Two types of representational theories generally distinguish the multiple approaches: discrete or dimensional. These are discussed in this section.

3.3.1 Discrete Representations

As a matter of course, there is a need to resort to language concepts for the theoretical and empirical research of emotion. The most common way to describe emotional states is by using categories derived from everyday language, such as anger, fear, and sadness. At first glance, it may seem obvious to use categorical terms to explain affective states because the terms used are so familiar, and it is natural to assume that such categories identify specific states. However, the literature suggests that discrete representations come with a range of complexities. One of the initial problems researchers are faced with is the vast number of available words in everyday language to describe emotion or emotional-related states. Cowie and Cornelius [16] pointed out several lists proposed by seven different investigations for the English language [168, 169, 170]. The size of these lists range from 107 to 558 words. Furthermore, lists that have been put forward on the web (see [30]), which are not specifically theory driven but rather reflect contemporary usage, amount to 3,000 words or standard phrases—280 of which occur in four or more sources. The sheer number of available categories raises problems for tractability. To complicate matters still further, the number of emotion words vary between different languages, certain words exist in some but not in others. This is not to say that the concept of a particular state described by a singular word in one language cannot be understood or described in another language without the respective counterpart. In this regard, many studies focus on similarities and differences in emotion descriptions that exist between various cultures and languages [171], while other studies specifically focus on the lexical semantics of emotion-related words [172].

Prototypical Categories

Researchers that emphasise categorical descriptions typically identify with a smaller set of *discrete* emotion episodes that have a special status of being relatively brief and fundamentally distinctive in nature. In discrete theories, the widespread assumption is that those privileged emotions are qualitatively unique, considered by most to be the clearest case of an emotional episode. Related terms such as ‘basic’ [36], ‘primary’, [173], ‘modal’ [174], ‘acute’ [175], ‘full-blown’ [16, 176], and ‘prototypical’ [177] are synonymously used for prime examples of

these emotions. These terms indicate a sense of hierarchy between emotional states. However, they can have various denotations depending on the theoretical context [16]. The term 'basic' used by Ekman and colleagues, for example, carries theoretical significance in Darwin's notion of evolution and its central role in the emotion phenomenon. On the other hand, for those who may describe emotions on dimensions (see section 3.3.2)—or otherwise—the term 'full-blown' is more commonly used [178, 179, 180, 156]. Cowie et al. [156] describe full-blown as states that have surpassed a certain limit of emotional strength within the Activation-Evaluation space. In this chapter, we will use the terms (interchangeably) in context with the relevant theoretical discussion. Thereafter, we will refer to full-blown (and underlying) emotions.

The Big N of Emotions

For many discrete theorists [173, 135, 36, 86], the idea that certain emotions are 'basic' is central to their research paradigm, mainly following Darwin's lead of evolutionary theory—assumptions also occasionally made in the Jamesian and Cognitive perspectives. In this regard, basic emotions should have evolved characteristics that should be functionally distinctive from other emotions, be universally recognised, be associated with hardwired neurological and physiological profiles, and should generate emotional response patterns that cohere and converge with each other. Although many share the view that some basic emotions exist, there is little agreement about how many there are and which ones those are. The different notions of basic emotions are associated with different empirical criteria. As Ortony and Turner [181] explained, "the divergence of opinion about the number of basic emotions is matched by the divergence of opinion about their identity". They distinguish between two criteria that characterise basic emotions, based on the conception of them being (1) biologically primitive or (2) psychologically primitive. While the perspective with regard to biological primitive emotions rests on the functional significance of evolution, the psychological primitive's perspective recognises basic emotions as irreducible constituents, which can account for all other complex emotions by a combination of these types. Although these proposed criteria commonly make up basic emotions, Ekman [182] has expressed his doubt with the idea that two basic emotions

can occur simultaneously to combine and form other compound emotions, such as 'smugness' being a blend of happiness and contempt. The second criterion is the assumption based on the evolutionary role that certain emotions must be readily apparent at birth, or early in life before learning has occurred. Although studies do indicate early signs of conveyed [86] and perceived emotion [183], it is difficult to determine which of these occur early enough to qualify as a basic emotion.

One of the most influential studies of basic emotion theory is that by Ekman. Based on facial behaviour observations, Ekman and colleagues [184, 185] accumulated a vast collection of evidence to support universality of facial expression across several different cultures. They initially suggested a list of six, which have become the most widely known list [77], termed by Cornelius as the "Big Six" [27]. These include "happiness", "sadness", "anger", "fear", "surprise", and "disgust". From a different theoretical approach, similar categories were suggested as being superior [1, 186], however, Ekman [182] later proposed a different list of 15 to include emotions not explicitly encoded in facial muscles. This list includes amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, relief, sadness/distress, satisfaction, sensory pleasure, shame, and pride in achievement. Because different commentators suggest different lists, some researcher now refer to them as the "Big" n emotions [179, 187], n being a figure greater than 2, and up to around 6 in the majority of cases [188]. Arguably, this highlights the potential uncertainty caused by the notion of the existence of basic emotions. According to Ekman, to qualify as a basic emotion, the characteristics that they share and that differentiate one from another, and from other affective phenomena, are based on the following criteria:

- Distinctive universal signals
- Emotion-specific physiology
- Automatic appraisal mechanism
- Distinctive universal antecedent events

- Distinctive developmental appearance
- Presence in other primates
- Quick onset
- Brief duration
- Unbidden occurrence
- Distinctive thoughts, memories, images
- Distinctive subjective experience

Several other attempts have been made to classify key emotions that reflect different perspectives, some of which are listed in table 3.1. The list of 'modal' emotions proposed by Banse and Scherer [189] is the most systematic in the literature specifically derived for speech research [16].

Subordinate Categories

Ekman considers all emotions to be basic, and does not acknowledge "non-basic" emotions [182]. This may raise the question regarding the variety of other emotions of which we are aware of. To explain his position, he introduces the concept of 'emotion families' that are similar in quality but may differ with respect to intensity [189], where *intensity* refers to the immensity of an emotional reaction, such as differences between rage and controlled anger. He argues that emotions are not single affective states but a family of related states. Each emotion family is endowed with a *theme* and *variation*. The theme of an emotion family shares the characteristics unique to a basic emotion, differentiated from one another by the criteria listed above. Themes are explained in evolutionary terms, while variation is explained by learning. Variations of a theme are the result of individual and contextual differences. This idea is familiarised by the work of Shaver et al. [1] whose work is based on prototype theory. His work does not explicitly identify basic emotions in terms of underlying biological substrates, but rather emphasises how people construct generic mental representations of important aspects

Lazarus (1999a)	Ekman (1999)	Buck (1999)	Lewis and Haviland (1993)	Banse and Scherer (1996)	Cowie et al. (1999b)	Plutchik (1980)	Parrot (2001)
Anger	Anger	Anger	Anger/hostility	Rage/hot anger	Angry	Angry	Anger
				Irritation			
				Cold anger			
Fright	Fear	Fear	Fear	Fear/terror	Afraid	Fear	Fear
Sadness	Sadness/distress	Sadness	Sadness	Sadness/dejection	Sad	Sadness	Sadness
				Grief/desperation			
Anxiety		Anxiety	Anxiety	Worry/anxiety	Worried		
Happiness	Sensory pleasure	Happiness	Happiness	Happiness	Happy	Joy	Joy
	Amusement		Humour	Elation	Amused		
	Satisfaction				Pleased		
	Contentment				Content		
		Interested			Interested		
		Curious					
		Surprised					
	Excitement				Excited		
		Bored		Boredom/indifference	Bored		
					Relaxed		
Disgust	Disgust	Disgust	Disgust	Disgust		Disgust	
	Contempt	Scorn		Contempt/scorn			
Pride	Pride	Pride	Pride				
			Arrogance				
Jealousy		Jealousy					
Envy		Envy					
Shame	Shame	Shame	Shame	Shame			
Guilt	Guilt	Guilt	Guilt	Guilt			
	Embarrassment		Embarrassment				
					Disapointed		
Relief	Relief						
Hope							
					Confident		
Gratitude							
Love			Love		Love		Love
					Affectionate		
Compassion		Pity					
		Moral rapture					
		Moral indignation					
Aesthetic							
						Trust/acceptance	
						Surprise	Surprise

Table 3.1: Recent list of key emotions, partly adopted from Cowie and Cornelius [16].

of emotions and the relationship between them. In their study, they compiled a list of several hundred English emotion terms and divided them into categories based on their similarity to one another. They analysed the data using hierarchical cluster techniques and found that English emotion words fall into 25 sub categories of synonyms. Six were identified as basic-level emotion categories (see Figure 3.1), namely "love", "joy", "surprise", "anger", "sadness", and "fear". They considered "surprise" questionable as it was less differentiated than the other categories. As we can see, this list is similar to the list initially suggested by Ekman.

Other investigations resemble the notion of hierarchal structures of emotions. With cognitive accounts of emotion, for example, appraisal components define an underlying emotion as a subset of the appraisal components of a full-blown emotion [71, 6]. However, cognitivists have a slightly different view of this. For example, rather than describing emotions as part of a hierarchal structure, one can conceptually describe underlying emotions as mixed emotions comparable to mixing a set of basic colors, such as the "Palette theory" of emotions [71].

Cover Classes

Full-blown expressions described by prototypical classes, such as the basic emotions mentioned earlier, tend to be most familiar to us, yet appear seldom in complex data that contain naturally occurring spontaneous emotions. Studies that choose prototypical categories often use elicited material that has been acted out, and are in most cases stereotypically exaggerated. When working with natural data, the emerging difficulties appear with the development of appropriate, restricted and manageable lists that describe regularly occurring states in everyday life. These states are commonly referred to as 'underlying' emotions. Having an unconstrained free-response format allows for detailed specificity and maximal accuracy of affect descriptions. However, it is difficult to quantitatively analyse free responses in a statistically robust way, as response occurrences for each label are generally too sparse [7]. If an unrestricted set of labels are used, researchers can reduce the number appropriately by merging them into a limited number of broad 'cover classes', often derived based on semantic resemblance [190, 191]. Alternatively, labels can be established in a data-driven way suited for a specific application.

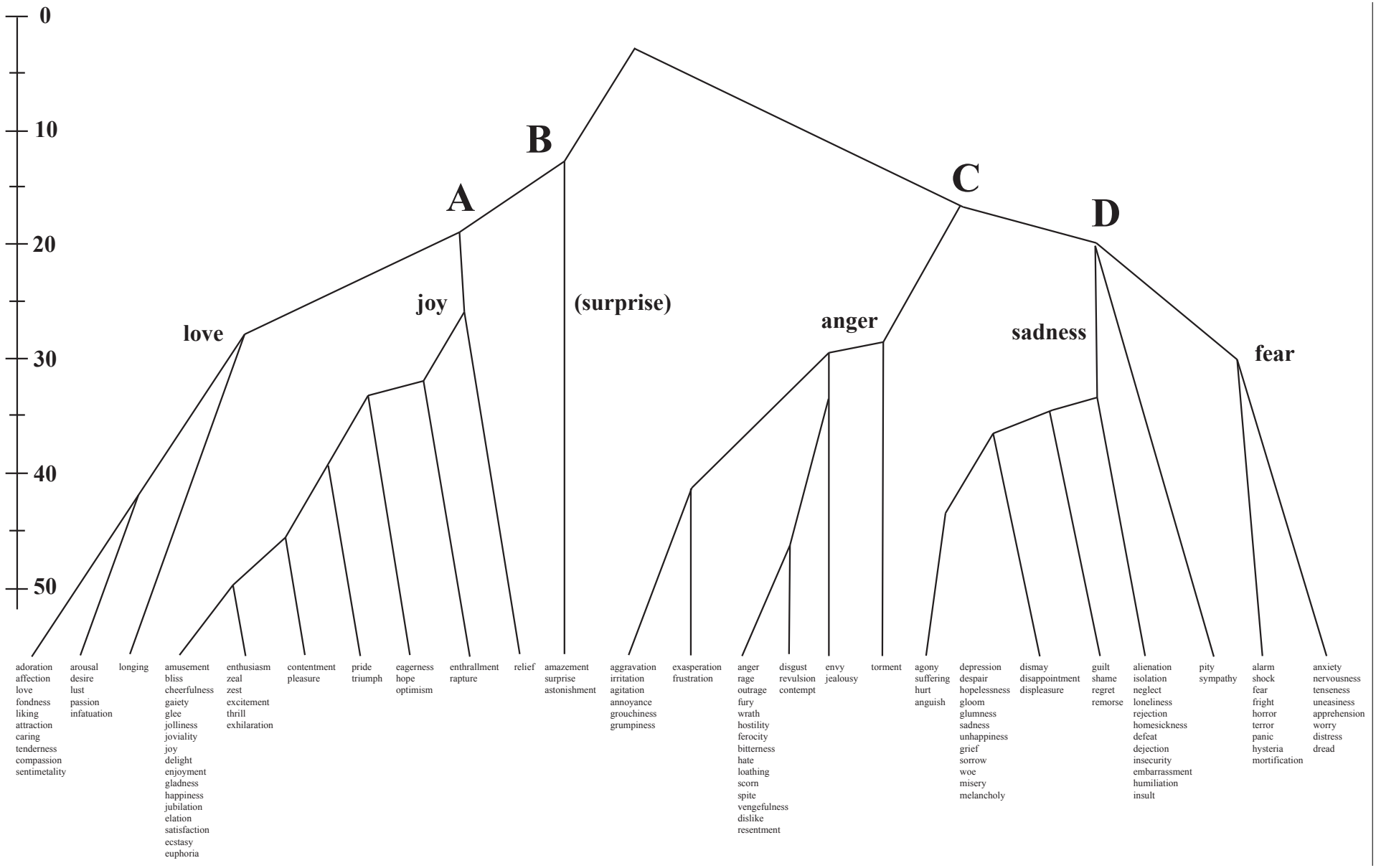


Figure 3.1: Emotion prototypes showing basic emotions with its relative underlying emotions [1].

That is, one may inspect the data and suggest labels by majority voting [192, 193], or in a manner in which labels seem to cluster in data sets [30]. However, effective methods for selection can be difficult, and the outcome of different cover classes limits comparability between studies. It is suggested that higher-order labels, such as basic or cover classes, are widely applicable, whereas more specialised labels are best suited for application-specific scenarios [194].

3.3.2 Dimensional Representations

Many researchers have suggested that categorical based descriptions pose too many difficulties, and as an alternative argue for dimensional models, a description well established within psychological literature [16]. Dimensional theories state that emotions can be described by an underlying characteristic and placed on coordinates along a given set of dimensions. Researchers can determine underlying structures that describe the order and diversity between emotion-related words or states. For example, participants may be asked to rate the similarity among pairs of words (semantic differential ratings), or asked to make introspective judgements about an emotional experience (verbal self-report) [195, 196]. By gathering this data, one can apply techniques such as *multidimensional scaling* and *factor analysis* to determine if any underlying dimensions are disclosed. It is often argued that dimensions give a more objective evaluation of emotional states. Jaworska and Chupetlovska-Anastasova [197], for example, state that emotion is a highly subjective and qualitative phenomenon, therefore the use of non-linear statistics, such as Multidimensional Scaling (MDS)[2] is particularly suitable.

Abstract Dimensions

Dimensional theory can be traced back as far as Wundt [199], who stated that bipolar scales represent "feeling opposites of dominating characters". Wundt proposed that human feelings, accessible through introspection, can be described by a position on three dimensions, namely

[2]Jaworska and Chupetlovska-Anastasova [197] indicate that the input data as qualitative is associated with non-metric MDS and input data as quantitative correlates with metric MDS. Holland [198] refers to the abbreviations MDS, NMDS and NMS, as Nonmetric multidimensional scaling.

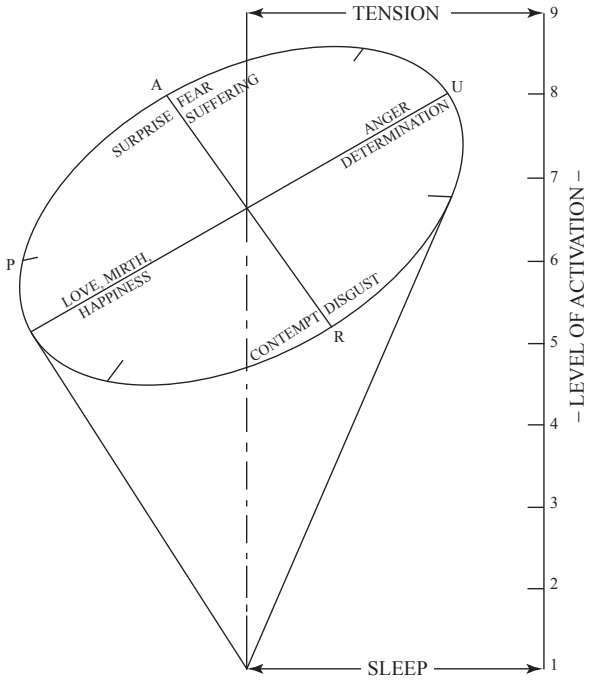


Figure 3.2: Schlosberg's three-dimensional model of emotion expression. [2].

"pleasure-displeasure", "excitation-depression" and "tension-release". Although his work did not seem to be based on accompanying empirical evidence, work thereafter by Schlosberg [200, 2] supported the concept on the basis of experimental psychology. In the context of rating facial expressions, Schlosberg proposed three dimensions of emotion: "pleasantness–unpleasantness", "attention–rejection" and "level of activation". The three-dimensional space is presented in Figure 3.2. He considered the level of activation to be quantitative in nature that could be coupled with concepts of physiological or neurological concepts. In contrast to discrete representations, he advises that dimensions allow rating divergence to be measured in a numerical form (cf. [201]), an aspect that is most certainly advantageous for scientific investigation.

Osgood et al. [202] carried out a semantic differential study to investigate underlying properties of affective structure in the English language. The major components underlying the meaning of natural language they concluded were the dimensions of *evaluation*, *potency*, and *activity (EPA model)*. In line with this, Averill [168] specifically studied the semantics of emo-

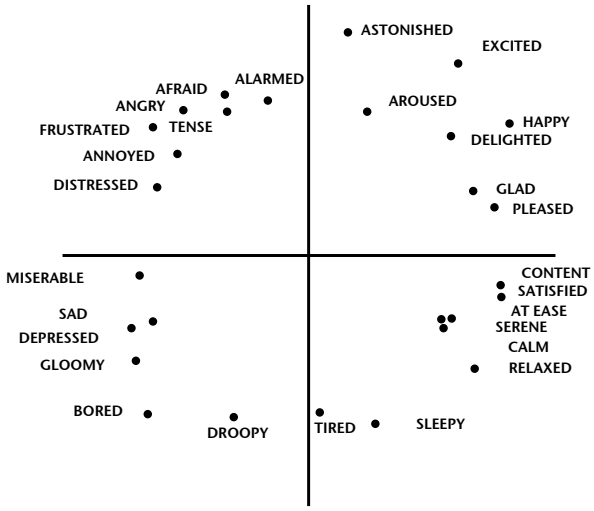


Figure 3.3: Russel's circumplex model of affect, with 28 emotion words on pleasure-displeasure (horizontal axis) and degree of arousal (vertical axis) [3].

tion terms and supported the evaluation and activity dimension. His interpretation of the data replaced the potency dimension with control and depth. Use of dimensions is supported by evidence based on intermodality responses, synaesthesia, physiological reactions, and semantic differential [87]. Albert Mehrabian and James Russell [203] suggested that there were three dimensions, namely *pleasure*, *arousal* and *dominance (PAD model)*. Russell and Mehrabian [204] found there was strong evidence to suggest the sufficiency of these three dimensions.

Russell [3] later reviewed the literature and presented an experiment using three different scaling techniques on the laymen's own mental representation of affective space, rather than an introspective account of their current state. The three different scaling techniques demonstrated a remarkable degree of consistency, from which they provided supportive evidence for the dimensions of pleasure and arousal. Although there was a commonly accepted distinction between data from introspective self-report and judgement data, he demonstrated that the findings of both sources of data exhibited similar structures. He, therefore, argued that the model is suitable for both layman's conceptualisation of affect and of affective experience. In other words, the model is not solely applicable to the perceived structure of emotion terms in natural language. Because the observed data is circular in nature (see Figure 3.3), the affective structure is referred to as the *circumplex model* of core affect (cf. [68]). He argues that dimensions are

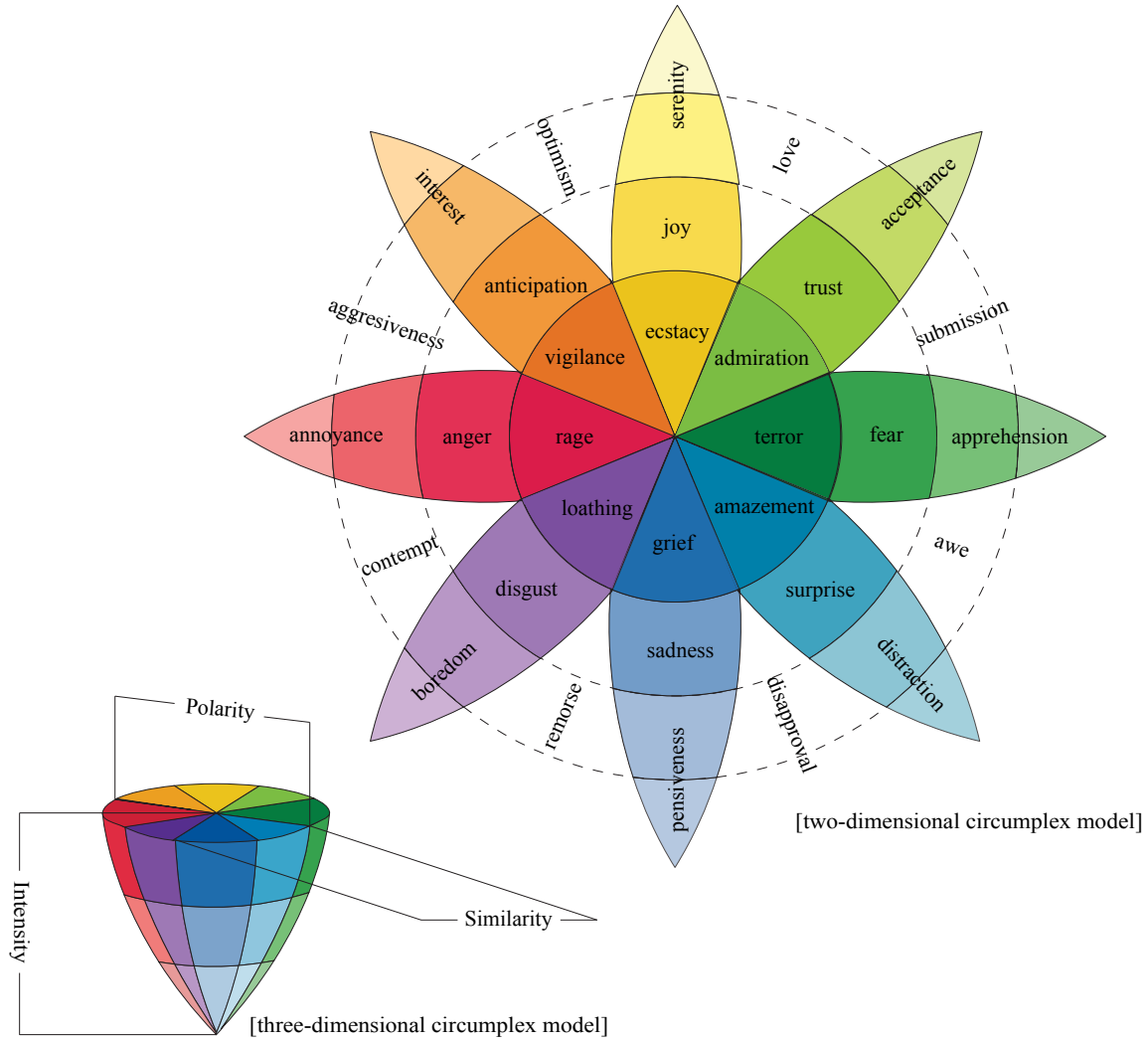


Figure 3.4: Plutchik's wheel of emotions partly adopted with labels for Intensity, Polarity, and Similarity, indicating the dimensions on the three-dimensional circumplex [4].

sufficient for adequately distinguishing between emotional states.

As mentioned, classification approaches in line with Darwin's theory of evolutionary emotion are traditionally affiliated with basic emotions. However, some researchers following Darwin's perspective also suggest the notion that emotions can be relatively conceptualised within some dimensional space. For example, Plutchik's work [34], based on psychoevolutionary theory, considers that eight 'primary' emotions can vary in similarity to one another, and can be conceptualised in terms of pairs of polar opposites. Within this space, emotions can also vary in intensity. Plutchik proposed both a two- and three-dimensional model, as illustrated in Figure 3.4. The three-dimensional model is presented as a cone. The vertical dimension represents intensity, and the inner circle represents degrees of similarity among emotions. Four pairs of

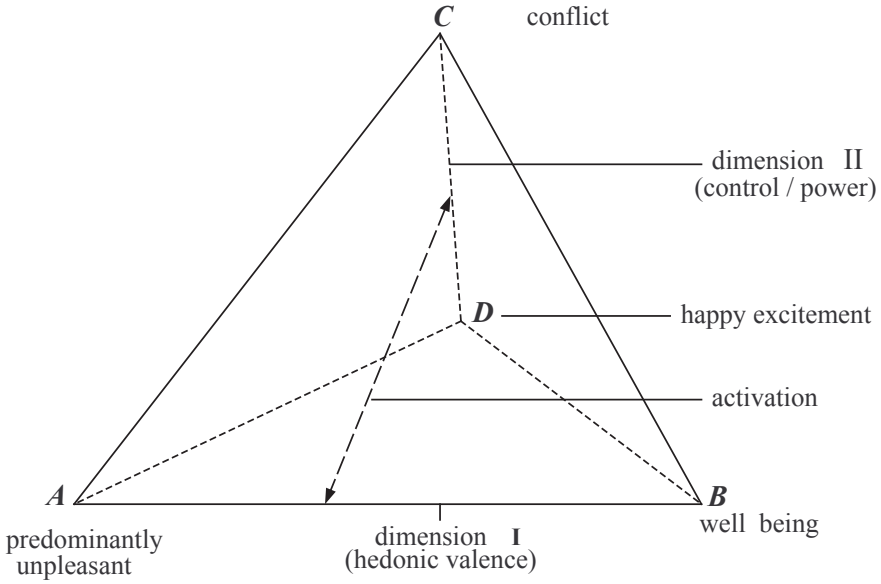


Figure 3.5: Scherer's Tetrahedral of hedonic valence, activation and control/power [5, p. 30].

opposites (polarity dimension) are depicted by eight categorical primary emotions. In the two-dimensional model, intensity is illustrated by sequential bands that decrease in strength towards the periphery, elaborated on by the different hues. Within the blank spaces, mixtures of two primary emotions are arranged. Plutchik proposes the analogy of primary emotions to basic colours, from which secondary emotions are formed by various blends of different primary emotions. The hue is used to express different intensities. This idea is similar to Scherer's "Palette theory" of emotions [71].

It is critically acknowledged [5, 158] that the interpretations made from the outcome of factor analysis and multidimensional scaling are subject to the nature of the stimuli used. In this regard, Gehm and Scherer state—referring to some of Russell's earlier work—that the outcome of these techniques can easily be biased by the partial selection of adjectives, so any one dimension can be strengthened or weakened by the inclusion (or exclusion) of a particular adjective. With this in mind, they include a large number of emotion adjectives in an attempt to cover emotion labels in their entirety. Furthermore, they examined the link between appraisal processes of the elicitation of emotion (component process model) with the dimensional structure of the semantic space of emotion words. The results from cluster analyses and MDS configuration

disclosed a tetrahedral arrangement (see Figure 3.5) representing three dimensions: "hedonic valence", "power/control", and "activation". Furthermore, much of the earlier research focused on prototypical types of stimuli, such as predefined lists for emotion words or stereotypical portrayals, whereas more recent research is concerned with realistic, spontaneous stimuli. As mentioned earlier, contemporary approaches, and their handling of naturalistic data, are often data-driven or application-specific. Taking this into consideration, a study by Batliner et al. [158] proposed dimensions that would be practicably suited for spontaneous speech (AIBO speech corpus), with a non-predefined set of labels. To do so, they conducted non-metrical multi-dimensional scaling (NMDS)[3] on the specified labels. By and large, categorical labels are nominal but can be modified to be non-dichotomous if mapped to dimensions. From their data, they interpreted two dimensions. As expected, the "valence" dimension is exhibited; however, instead of Arousal, they suggest an 'interaction' dimension.

Based on semantics, memory recall, and facial and vocal expression, the empirical evidence demonstrates that emotion-related states can be adequately mapped onto a two-dimensional or three-dimensional space, and subsequently be distinguished according to their position [196]. Although different interpretations are made about the underlying dimensions, mostly, the elemental two-dimensions proposed are conceptually similar. On a two-dimensional model (Evaluation/Valence and Activation/Arousal), anger and fear are often placed very near to each other (see Figure 3.3), since both have strong negative valence and high arousal. Intuitively, however, they seem clearly different on a subjective and behavioural level. A third dimension, such as Potency, Control, or Dominance, has been used where the description of a two-dimensional model falls short to this effect [152, 24, 207]. Moreover, the ambiguity of "surprise" has long been discussed in the literature [208, 1]. It is explained that "surprise" has an ambiguous valence, as it can be associated with either positive or negative emotions [24], and it can occur simultaneously with other emotions [208]. To address this, a fourth dimension characterised by appraisals of novelty and unpredictability has been suggested to distinguish the state of "sur-

[3]NMDS is a scaling method that disregards the properties of distance between certain quantities measured [205, 206]. It retains the similarities (or dissimilarities) between values, by preserving the rank order but not the exact numerical values.

prise" from other emotions [208]. Certain "surprise" states may not be adequately described by a single-word label, and instead two emotion words are needed, such as "joyful surprise". Although it seems slightly paradoxical, one could argue that dimensions demonstrate a higher degree of qualitative differentiation in such instances.

In summary, it is evident that different inferences are made about the emerging factors (illustrated in Table 3.2.). The naming of dimensions by the individual researchers are based on their interpretation of the data, therefore the terminology used for dimensions is somewhat equivocal [32]. In other words, ambiguities can be expected with factor identification. In fact, Scherer [170] acknowledges that factor analysis and multidimensional scaling are open to criticism for this reason. There are three dimensions that are most prevalent. Synonymous terminology has been used but they are conceptually similar [32, 157]. These are as follows:

1. **Evaluation, Valence, or Pleasure:** the most important element widely agreed on is that of the hedonic valence of an emotion. It is concerned with the positive or negative evaluation of people, things, or events [156]. In this regard, the hedonic valence is assessed focusing on the individual's subjective feeling state, or the cognitive appraisal of external stimuli.

2. **Activation or Arousal**: the second most agreed element is the activity/arousal factor. This measure represents the degree of alertness, excitement, and the organism's disposition to engage in action [70]. This dimension is often associated with the physiological arousal and neurological activation.

3. **Control, Potency, or Power:** This dimension has a more controversial history in emotion research [207]. It refers to the sense of power or control a subject has over the eliciting event.

Appraisal Dimensions

Central to the cognitive perspective is the cognitive evaluation made by an organism of an emotion-relevant object, event, or situation in its environment. Every emotion correlates to a

Researchers	Assessment	Abstract dimensions			
Wundt (1874) [199]	Introspection (theory)	pleasure-displeasure	excitation-quiescence	tension-relaxation	
Schlosberg (1941) [200]	Judgements of facial expressions	pleasantness-unpleasantness	attention-rejection		
Osgood et al. (1957) [202]	judgements of meaning in natural language (semantic differential)	evaluation	activity	potency	
Osgood et al. (1966) [209]	Judgements of facial expressions (factor analysis)	pleasantness	activation	control	
Bush (1973) [210]	Similarity judgements of adjectives (MDS)	pleasantness-unpleasantness	level of activation	level of aggression	
Mehrabian and Russell (1974) [203]	Self-reports of imagined events (regression analysis)	pleasure	arousal	dominance	
Green and Cliff (1975) [211]	Judgements of emotional speech stimuli (NMDS, factor analysis)	pleasant-unpleasant	excitement	yielding-resisting	
Russell (1980) [3]	Review of literature & similarity judgement of emotion words (MDS, UDS, PCA)	pleasure	arousal		
Plutchik (1980) [34]	Similarity judgements of emotion words (semantic differential)	intensity	similarity	polarity	
Scherer (1984) [170]	Similarity judgements of emotion words (cluster analysis, MDS)	positive-negative (evaluation)	activity	potency	
Gehm and Scherer (1988) [5]	Similarity judgements of emotion words (MDS, cluster analysis)	valence	activation	control/power	
Watson et al. (1985) [212]	Similarity judgements of emotion words (NMDS) studies	positive affect	negative affect		
Shaver et al. (1987) [1]	Memory tasks of emotion pictures and sounds	evaluation	activity	potency	
Feldman (1995) [213]	Similarity judgement of semantics and self report of mood (factor analysis)	pleasantness	arousal	dominance	
Church et al. (1998) [214]	Cross-cultural similarity judgements of emotion-words (hierarchal cluster analysis, MDS)	pleasant	arousal	certainty-uncertainty	
Barrett (1998) [215]	Self report of affective states (factor analysis	valence	arousal		
Batliner et al. (2007) [158]	Judgements of spontaneous emotion speech data (non-metrical multi-dimensional scaling)	valence	interaction		
Fontaine et al. (2007) [208]	Cross-cultural study of the semantics of emotion terms using a componential approach (PCA, GRID instrument [7])	evaluation-pleasantness	activation-arousal	potency-control	novelty/unpredictability

Table 3.2: Suggested dimensions by different commentators.

unique appraisal of a situation, which changes according to appraisal variations [216]. Ortony et al. [74, 6] designed a cognitive emotion model, called the OCC (Ortony, Clore and Collins) model, which is implemented for the benefit of computational modelling [217, 87, 218]—valenced reactions contingent on conditional rules. In this model, appraisals of stimuli distinguish between 22 emotion types. As illustrated in Figure 3.6, the uppercase labels represent structural elements, whereas lower case labels correspond to potential emotional states. The model has three main branches that represent valenced reactions to three types of stimuli: *consequences of events*, *actions of agents* and *aspects of objects*. The first branch represents the emotions associated with reactions to an event and its consequences, which one can appraise as a *pleasing* or *displeasing* affective reaction. The first distinction is illustrated according to whether the consequence of the event concerns oneself or of others. The *consequences for other* of the event can be *desirable* or *undesirable*, and depends on how one feels towards the other. On the one hand, if the consequence of the event is desirable for the other, one can either feel *happy-for* or *resentment* towards the other. On the other hand, if the event is undesirable for the other, one can gloat at the other's misfortune or feel pity. On the other sub-branch, if the potential consequence of the event is about oneself, one appraises the prospects as *relevant* or *irrelevant*. In the case of a relevant prospect, the potential emotional reaction will be *hope* or *fear*, for which the anticipated prospect is evaluated as *confirmed* or *disconfirmed*. For example, if one hopes for something to happen, and it is confirmed, one will feel *satisfaction*. This group of emotions is called *prospect-based* emotions. If the prospects of an event are irrelevant, the elicited emotion will be either *joy* or *distress*, depending on the appraisal of the consequences of the event. If the focus is on the *self*, the elicited emotion may vary between *pride* and *shame*, whereas if the focus is on the *other agent* one may feel *admiration* or *reproach*.

In relation to the middle branch, all emotions correspond to an *approval* or *disapproval* of the *actions of an agent*. The differentiation made within the *attribution* group of emotions depends on the agent's focus, whether the agent is the self or other. When we react to both an event and an action, the *well-being* and *attribution* type emotions are combined to form the *well-being/attribution compounds* group. The emotion *gratification*, for example, is the combination

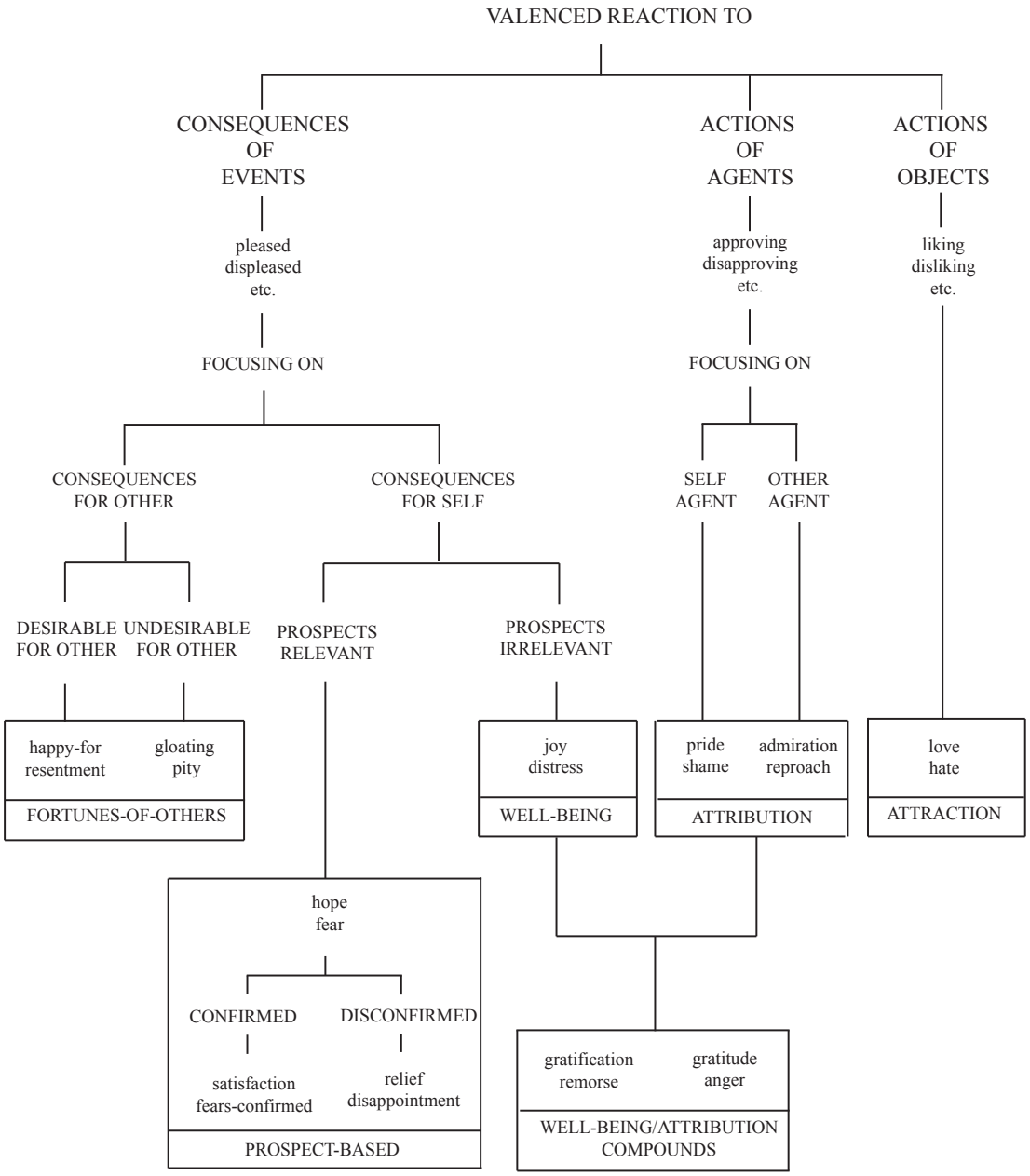


Figure 3.6: The OCC model: global structure of emotion types [6]. The above illustrates valenced reactions to three types of stimuli: consequences of events, actions of agents and aspects of objects

of *joy* (reaction of an event) with *pride* (reaction to an action). The last branch corresponds to elicited emotions that are reactions to aspects of objects. These emotions correspond to a *liking* or *disliking* of objects or aspects of objects. This group is called the *attraction* type of emotions.

A well-developed alternative model is that of Scherer's [219, 170], the *component process model*. In this model, differentiated emotions are the results of successive outcomes of appraisal

described by a number of *stimulus evaluation checks* (SECs), along different dimensions. The following is a brief list of the involved sequences (see [219] for further details):

- **Relevance** includes checks for novelty, intrinsic pleasantness, and goal/need relevance.
- **Implications** include checks for causal attribution, outcome probability, discrepancy from expectation, goal/need conduciveness, and urgency.
- **Coping potentials** include checks for control and power of the event.
- **Normative significance** includes checks for internal standards, and external standards.

Scherer [170] conducted a similarity judgement experiment using cluster-analytic and multi-dimensional scaling techniques on natural language labels. As a result of the study, Scherer supports that appraisal criteria of the component process model can be organised within the semantic space of Evaluation and Activation, thus, linking appraisal theory to dimensional models of affect. More recently, he stated [7] that numerous studies show considerable influence on emotion differentiation by appraisal dimensions, notably goal conduciveness (conducive/obstructive) and coping potential (control/power). Figure 3.7 shows a mapping of Russell's [220] proposed circumplex, with valence and activity as the dimensions, and Scherer's [170] own results organised by appraisal criteria—the results were superimposed rotating the axes by 45 degrees. Another study [196] explored the three-dimensional space of affect (valence, arousal, potency) to examine the position of elicited feeling states, again based on the results of appraisal profiles. However, they acknowledged a limitation of their study. They recognised that the use of picture stimuli is not ideal for appraisal criteria—examining an individual's goals and needs would be somewhat restricted—therefore, limiting comparisons between two- and three-dimensional models. The results, nevertheless, suggested that goal conduciveness correlated highly with the appraisal of pleasantness, which is specified by valence. The arousal dimension appeared to be regulated by novelty and unexpectedness, and the control/power dimension is determined by the appraisal of coping potential. Meanwhile, a pilot study within HUMAINE research showed that these appraisal variables—that appear to be linked to the three aforementioned dimensions—tend to receive high agreement with labelling,

Figure 3.7: Russell's results of a two-dimensional valence by activity space superimposed on Scherer's results based on similarity ratings of emotion terms [7]

namely that of *intrinsic pleasantness, goal conduciveness, expectedness, power/powerlessness of event and consequences* [221], while the majority of appraisal descriptors received low rater agreement.

Based on the component process model, Banse and Scherer [189] confirmed detailed physiological predictions about the vocal changes associated with different emotions. Using acted speech material, they demonstrated that component patterning, as a consequence of sequential event appraisal—or stimulus evaluation checks—predicted the configuration of vocal changes for happiness, sadness, cold anger, and panic fear. Predictions deviated, nevertheless, for contempt, anxiety, shame, disgust, hot anger, and elation.

When deciding if appraisal-based representations are suited for a particular study, there are two distinct descriptions that a researcher should consider: (1) researchers may be concerned either with what the speaker is truly feeling and identify the state that caused the emotions, or (2) researchers may be concerned with the impression the expressed emotion has on a typical listener. As mentioned previously (section 3.2.5), this distinction between emotion descriptions has been termed as 'cause-type' and 'effect-type', respectively [156]. For cause-type studies, one can operationally define the elicited emotion-related state through the description of the objective eliciting methods. For instance, a study by Fernandez and Picard examined drivers' speech under stress [153], whereby the stress level was operationalised in the context of varying conditions of cognitive load. Similarly, appraisal theories allow researchers to monitor detailed specifications of appraisal dimensions in the inducing procedure of an emotion. In the framework of Scherer's component process model this is done by systematically manipulating sequential cognitive evaluation checks of emotion-antecedent events (novelty, pleasantness, goal relevance, etc.) [176, 222]. Intrinsically pleasant and unpleasant images, for example, may be used to manipulate the appraisal criteria of pleasantness, while a specific event in game play can be used to manipulate goal conduciveness [223].

3.3.3 Resources

In this section, we review some of the key instruments and resources used to assess observed behavioural or self-reported affective states. The selection of tools described here are by no means exhaustive, but they provide us with guidance for our specific goals. The tools we describe have been used previously for emotion in speech research, except measurements using FACS, as this is used to measure facial expressions. Because studies of face behaviour are long-lived, we include the FACS tool in our review to demonstrate the extent to which observable behaviour can be systematically assessed. We first introduce the FACS tool, and then demonstrate instruments that embody both discrete and dimensional representations.

Facial Action Coding System (FACS)

The Facial Coding System (FACS, [85]) measures observable action (contraction and relaxation) of isolated facial muscle movements—called Action Units (AUs). This system requires high levels of expertise with up to 100 hours of training. By itself, it is an index system for facial expression, and describes facial movements without implications about the behaviour itself. Instead, the Emotion Facial Action Coding System (EMFACS) [224] uses FACS scoring just on facial behaviour that may have connotations of emotion. It is an abbreviated version of FACS and only uses a selective subset of the FACS coding system. However, in this system, facial actions are believed to be associated with certain discrete expressions, such as "happiness", "sadness", "surprise", etc. The authors state that it is unsuitable for disguised, highly controlled, subtle or blends of emotions[4]. FACS is a descriptive analysis of behaviour, rather than an inferential process. Nevertheless, together with empirical findings of experience or inference, FACS can be used to be associated with scoring criteria for expressions of certain discrete emotions—as Ekman's work demonstrated—or emotion dimensions [225].

It has been shown that facial and acoustic features are strongly interrelated. Busso and Narayanan [226] showed that the relationship between facial gestures and speech is influenced by emotional content. They suggest that their results are beneficial for applications such as facial animation and multimodal emotion recognition.

Emotion Lists

In section 3.3.1 we mentioned that theoretically derived lists can range from 107 to 558 words for non-basic emotions. Cowie et al. [30] provide a variety of website links that list emotion words and phrases that reflect everyday usage, rather than categories chosen from theoretical assumptions. To give an example of the vast amount of words that are used to describe feelings in natural language, one of the sites[5] lists 4000 words and provides over 600 words to describe negative feelings alone. Cowie et al. inspected the sources and found that 280 occur in four or

[4] see http://www.face-and-emotion.com/dataface/facs/emfacs_intro_authors.html for EMFACS details of use.
[5] `http://eqi.org/fw.htm`.

more sources. They reduced the list to 43 cover classes (see Table 1, [30, p. 226]) that they found were most observed in naturalistic data deduced from several studies [227, 228, 221].

The Grid Instrument

The GRID[6] instrument was developed by Scherer and colleagues [7, 208] to address the semantics of emotion terms using a componential approach. This instrument consists of a questionnaire that gathers data to assess the meaning of emotion words across 20 different languages. The questionnaire comprises 24 emotion terms and 144 emotion features that represent different components of emotion (e.g. appraisals, bodily reactions, expressions, action tendencies, and feelings). So far, empirical evidence demonstrates that four dimensions can adequately describe the semantic space covered by the emotion terms. These are 'Valence', 'Power', 'Arousal', and 'Novelty'.

Positive and Negative Affect Scale (PANAS)

The Positive and Negative Affect Scale (PANAS) is a tool that consists of two 10 adjective mood scales to measure affective states [229]. PANAS consists of discrete categories subdivided into positive (attentive, interested, alert, excited, enthusiastic, inspired, proud, determined, strong and active) and negative (distressed, upset, hostile, irritable, scared, afraid, ashamed, guilty, nervous and jittery) categories. Each term listed is then rated using a five-point scale indicating how much the emotion is present. In some respect, this tool is similar to other valence type models.

The Profile of Nonverbal Sensitivity (PONS)

The profile of nonverbal sensitivity (PONS), developed by Rosenthal, Hall, DiMatteo, Rogers and Archer [143], consists of a scale to investigate how humans infer nonverbal information via facial, body and vocal cues (see [230]). PONS is composed of discrete auditory and visual information portrayed through individual segments, both presented in a randomised matter.

[6]For more information refer to: http://www.affective-sciences.org/grid.

The visual information is displayed in three types of cues: body, face, and a combination of both (figure cues). The auditory information is presented as a spliced voice segment (RS) or a content-filtered voice (CF). The different combinations of visual and auditory cues are referred to as "channels". The relevance of voice filtering was used to remove the lexical aspect of the message, leaving the "tone" of voice[7] present.

Diagnostic Analysis of Nonverbal Accuracy (DANVA)

The Diagnostic Analysis of Nonverbal Accuracy (DANVA) tool, designed by Nowicki and Duke [231] is an instrument designed to evaluate ability to accurately process nonverbal information. It assesses facial expression, posture, gesture and "tone" of voice. The test for the perception of facial expressions uses images of happy, sad, angry, fearful, or neutral expressions. If the participant feels the emotion is not listed, the choice would be classified as "other". The test on the perception of tone of voice involves a semantically neutral sentence spoken to portray each of the four emotions (happy, sad, angry or fearful). To test the expressive component, participants were asked to speak a semantically neutral sentence after being described a situation to them designed to elicit one of the four emotions. Good internal consistency (.68 to .88) and a good test-retest reliability (.70 to .86) have been demonstrated for this tool [230].

Self-Assessment Manikin Test (SAM)

The dimensions Pleasure, Arousal, and Dominance (PAD) have been used as a Semantic Differential scale [204, 232]. The Self-Assessment Manikin (SAM) is a graphical depiction of the PAD emotional model (Figure 3.8). This easy to use rating tool was developed by Lang [8] as an alternative to verbal emotion assessment, making SAM language-free and suitable for cross-cultural studies [233]. The three dimensions are represented in the form of graphical characters from which the user can choose an emotional state. This tool has been used by many researchers for assessing emotion [234, 235, 152, 23, 236, 98, 237]. Morris et al. [238] compared their results with the results obtained in the Mehrabian and Russell [204] study, claiming

[7]Westerman et al. [230] refer "tone" of voice to non-lexical information transmitted with verbal messages.

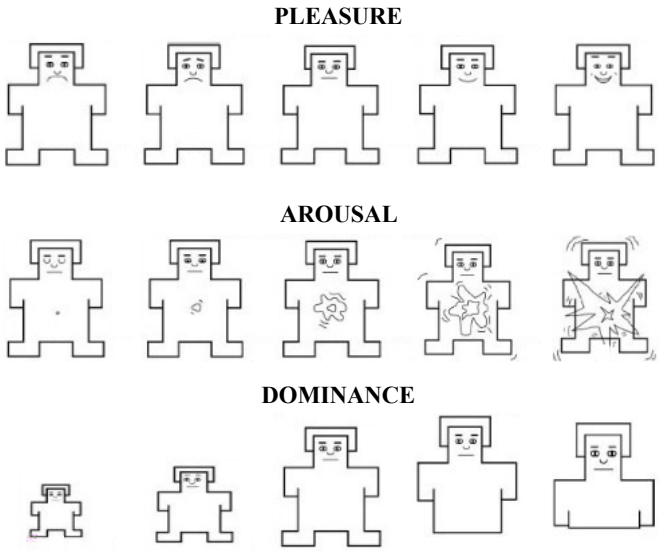


Figure 3.8: Self-Assessment Manikin test (SAM) [8]. In most cases, the labels "Pleasure", "Arousal" and "Dominance" are not shown to the participant during the experimental procedure.

sufficient similarity. In their study, a combination of SAM and a set of emotion adjectives were used to analyse emotional responses to television commercials, and showed that measurements were comparable to measurements across various groups.

Trace Tools: Temporal Measurement

Cowie et al. [9] developed a tool to track emotional content as it changes over time, called the *Feeltrace* tool (Figure 3.9). It has two emotion dimensions, *Activation* (from passive to active) and *Evaluation* (from negative to positive). In this system they use a colour coding system derived by Plutchik [34]. They found the tool comparable to using an emotion vocabulary of 20 words or more. While they recognised the limitations of the tool, they listed several advantageous. The tool can handle in-between states, it comes with statistical advantages because of obtained numerical data, and it has the unique ability to measure temporal variations, which seems particularly suitable for speech. However, there is a lack of ability to capture details that distinguish certain emotions, particularly the often-cited example of fear and anger. Understandably, reducing emotion to two dimensions may lose some detail for certain states. In this regard, Grimm et al. [152] found the tool was inadequate for their study because of its

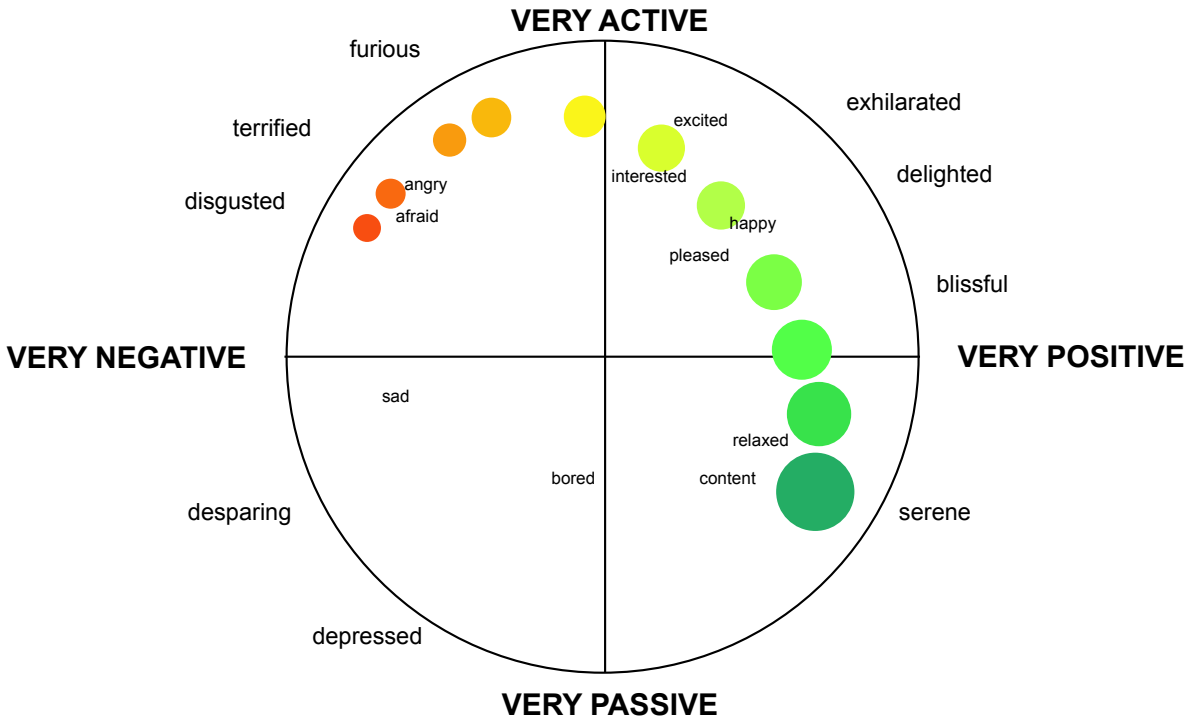


Figure 3.9: The Feeltrace two dimensional labelling tool that allows measurement of an emotional state continuously over time [9].

restrictions on a two-dimensional space. Furthermore, they argued that valence is a bipolar rather than an angular-type periodic entity. To distinguish between emotions such as fear and anger, they favoured the three-dimensional PAD model (i.e. the SAM tool, Figure 3.8).

In a similar vein, tracking is accessible on emotional intensity, using the *ETraceScale* (see Figure 3.10). Emotional intensity can be quantified by moving the mouse to follow perceived differences. Intensity is represented by different colours, ranging from blue (zero intensity) to red (maximum intensity), and gradual changes are represented by circle size. The scale is labelled at interval markers to indicate to the rater the different emotion intensities: zero emotion, mild social emotion and emotion at maximum intensity. The *EtraceCat* (Figure 3.11) is a counterpart of EtraceScale that quantises motion into categories, rather than having continuous tracking. The categories are 'completely emotionless', 'partial emotion', and 'emotion in the full sense'.

A comprehensive use of trace-type labelling is included in the Humaine database [194], a naturalistic database of multiple modalities. For this database, they extended the use to include:

- the intensity of the emotion (IntensTrace)

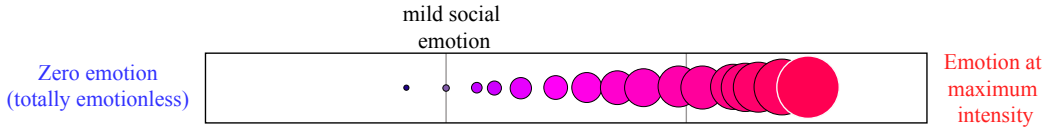


Figure 3.10: ETraceScale allows for real-time labelling of Intensity relative to time, developed by Cowie et al. (diagram obtained from Steidl [10]). Raters are able to use the mouse, pressed down, to record emotion intensity from moment to moment.

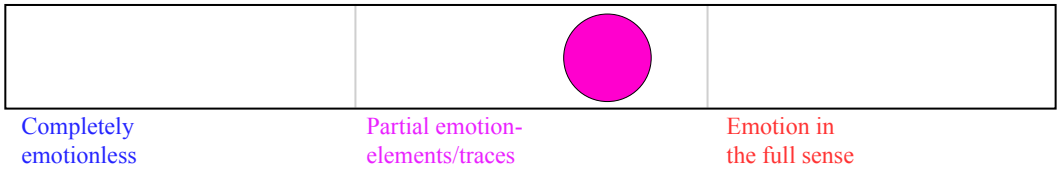


Figure 3.11: ETraceCat demonstrates the intensity as categories, developed by Cowie et al. (diagram obtained from Steidl [10]). Raters are able to use the mouse, pressed down, to record emotion intensity from moment to moment.

- the level of which the emotion appeared genuine or acted (ActTrace)
- the extent to which a person is trying to cover the emotion actually felt (MaskTrace)
- the dimensions of power (PowerTrace) and unexpectedness (Anticipate/ ExpectTrace)
- the intensity of the highest ranked emotion word for a clip (WordTrace), such as fear.

Geneva Emotion Wheel (GEW)

The study mentioned earlier by Scherer (section 3.3.2) attempted to amalgamate the scientific concepts of emotion with layman concepts of emotion. By mapping the results derived from the component process model onto Russell's circumplex model, he examined the link between appraisal profiles and the dimensions often used in self-reports studies, Evaluation and Activation. In his work, he created a graphical representation of emotion family members, whereby the emotion intensity is presented by circle size, being more intense at the periphery than at the centre. The choice of concrete families were inspired by what are generally considered 'basic' or 'fundamental' emotions. Rather than a smaller set of basic emotions, a total of 16 emotion families were chosen to facilitate ease of reading. This prototype, called the *Geneva Emotion Wheel* (GEW) is shown in Fig 3.12.

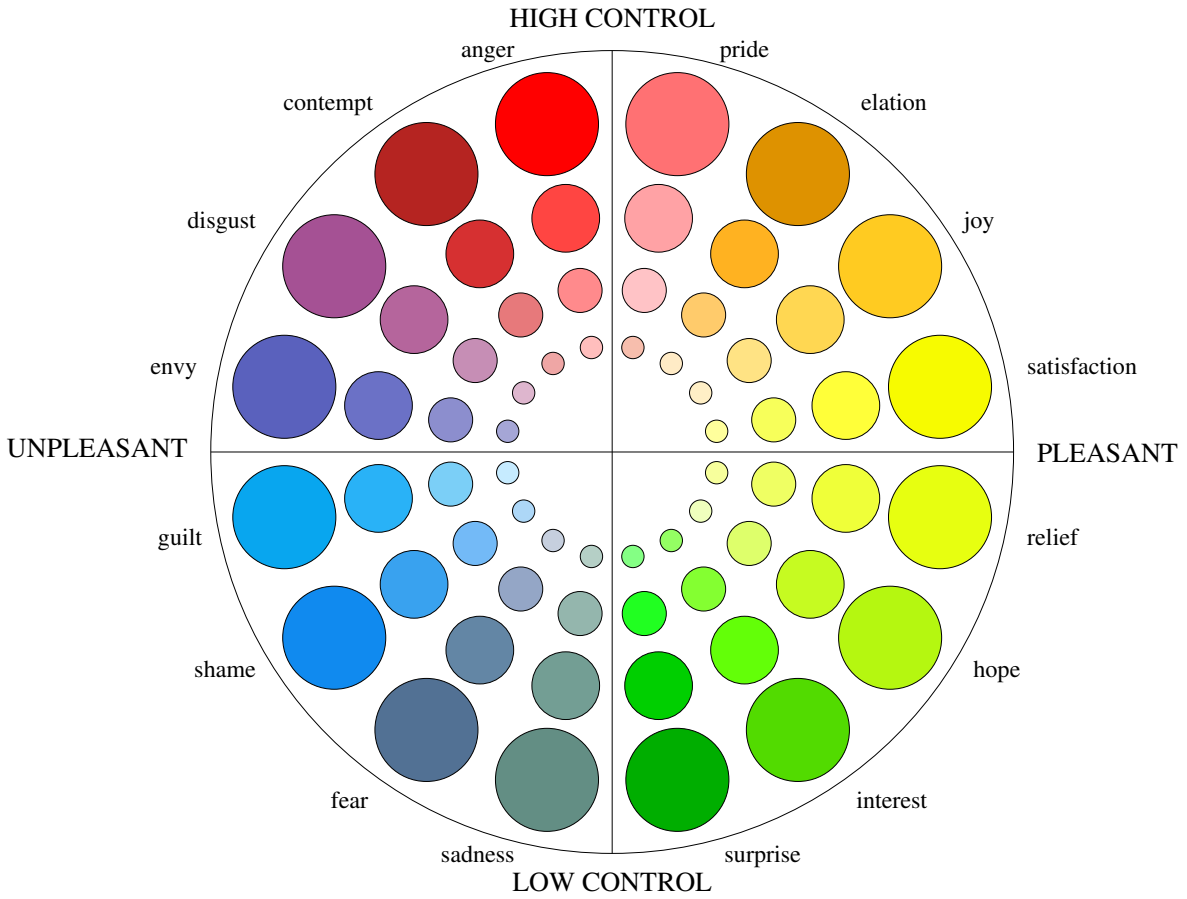


Figure 3.12: Geneva Emotion Wheel (GEW) demonstrates a dimensional model that incorporates discrete labels and emotion families [11].

This instrument is used to measure the affective responses of participants to a variety of stimuli. These can include objects, situations and events. The emotion families are arranged around the wheel in accordance to appraisal dimensions, namely as Control/Power and Pleasantness-Unpleasantness. Participants are informed that the represented words for each family are simply an indication to a range of closely related emotions. The participants are able to choose from two emotion families if a blend seems to be present at different intensities.

The GEW conveys potential from both a discrete and dimensional approach. The dimensional aspect adheres to the comparability of emotions. The visual representation of intensities and the categories assigned to emotion families allow for a conventional association of emotion. The GEW is a relatively recent development, and more testing needs to be done for the validity of the tool before it is established for general use. It has, however, been given some use in emotion-related research [239, 22].

3.3.4 Labelling Naturalistic Emotional Speech

The development of naturalistic speech databases (see [240, 241, 242, 201] for a list of existing corpora) revealed several problems with annotating that are not present with acted material. An early example of a naturalistic emotional database is the Leeds-Reading database [243]. This database contains real-life emotional situations with an emphasis on intense material. Speech was extracted from interviews with a psychologist and from broadcast material. Douglas-Cowie et al. [201] provided a review of a series of three studies conducted on the material. Emotional annotation was organised into four levels. The first level used freely chosen everyday emotion labels. The second level specified the strength of the emotion, together with a sign to indicate valence. The third and fourth level described emotional episodes based on the individual's appraisal of the event. They specified that the number of categories associated with an in-depth qualitative coding strategy would amount to smaller occurrences in each category. Although broadcasts provide emotionally rich material, this corpus is not publicly available due to copyright issues, as is the case for most truly natural material.

The development of the Belfast Naturalistic database [244] followed from the Leeds-Reading database and overlapped in descriptive schemes, exploring the audio-visual data from the broadcast material. Their focus was to develop a quantitative description, and address two aspects that they concluded marked naturalistic data: uncertainty and gradation. To deal with uncertainty, the database included all individual ratings rather than solely determining agreement. To address gradation, they developed *trace* techniques to evaluate and quantify emotion as it changes over time along underlying affect dimensions: Evaluation and Activation. For the rating task, they acquired three trained raters to use the tool. They argued that using quantitative measurements, with a tool such as *Feeltrace*, better estimated agreement levels because labels can be evaluated as similar or dissimilar [201]. As somewhat unexpected, dimensional rating showed less individual differences, with agreement being closer on the evaluation dimension. To measure rater agreement on time-continuous evaluation, the mean of the cursor position represented the selected categories. As an alternative to averaging cursor positions, the scale

can instead be discretised into a finite number of categories. Although, discretising continuous dimensions may be better suited for discrete (quantised) speech segments.

According to Cowie et al. [30], the most extensive collection of material to date is found in the JST/CREST Expressive Speech Corpus [245]. It contains several sources of material for analysis, such as television broadcast, DVD and video, but its main emphasis is the collection of volunteers who recorded conversations in everyday situations wearing sound recording equipment. The schema for labelling included the *Feeltrace* tool (as above). They note that labellers understood the meaning and validity of the two dimensions, Valence/Arousal[8]. The work on this concluded that the chosen framework was not sufficient to describe the expressions found in the material. Therefore, they proposed three levels for labelling: state of speaker, style of speaker, and physical aspects of the voice—each level contained up to five further subcategories. This comprehensive schema appears to be necessary when listening to speech in context and over long segments. For the task of labelling, they familiarised themselves with the speaker's mannerism from the material obtained over a five-year period. Hence, this descriptive scheme is intuitively data-driven. For this thesis, where short segments of speech are rated, such a comprehensive scheme may not be suitable to the same degree.

In most naturalistic speech datasets, clearly pronounced emotions, such as the full-blown prototypical types are not regularly conveyed. While most labelling schemes for acted data contain a predefined list of well-established emotion categories, these categories do not generally apply to spontaneous speech that occurs in daily situations. In many cases, ad hoc labels that are appropriate in the context of a particular research—application-dependent or data-driven—are carefully chosen. Unfortunately, choosing a descriptive framework for naturalistic emotional speech remains a difficult task, and many things need to be considered. Cowie et al. [30] summarise some of the issues surrounding labelling databases:

[8]Commentators recognise some ambiguity between dimensions used—we use the terms Evaluation/Activation instead of Valence/Arousal

- *Economy.* The vast number of available words in everyday language to identify emotional states makes it an unwieldy descriptive system. To address this, fewer primitive categories can be chosen based on theoretical selection.

- *Consistency.* As mentioned in section 3.3.1, broad 'cover classes' are often proposed to achieve economy. However, cover classes are data-driven, and different datasets, therefore, suggest different cover classes. It is suggested to base cover classes on theory to reduce the chance of varying lists.

- *Intermediates.* Naturalistic emotional databases appear to contain elusive states that are not sufficiently described by a singular every-day term. To overcome this, abstract or appraisal-related dimensions are suggested.

- *Semantics.* The capacity of semantics being associated with a label is an important aspect for use in a functional system, where not only the state is recognised but also what caused the state to arise. Appraisal-based labelling offers ways of representing meaning with a label.

- *Natural classes.* Although one can suggest that choosing classes based on theory should be less problematic, it is difficult to know whether well-chosen classes can be refined to suit all envisaged applications. Classification is often application-dependent, and choosing suitable classes for the right application is a subtle problem.

3.4 Discussion

Assessments provide the labels that are necessary to identify emotions. Because emotions are pervasive in a number of response systems, emotion can be assessed in several ways. In fact, we have already noted that emotion is commonly accepted as consisting of several components. In this chapter, we first discovered that labels provided by self-reports, physiological measurements, and behavioural observations are established corresponding to one of these components (section 3.2). The investigation of emotion involves linking the results of one type of measurement with another. For example, physiological measures (including acoustical data) are

linked to either the (subjective) experiential component or the (observable) behavioural component, or both. Although self-reports are the only means to obtain information about the subjective experiential component of emotion, measuring judgements of emotion (behavioural component) offer better validation methods, and are therefore considered more reliable (section 3.2.3 and 3.2.5). Although distinct in nature, some theorise that all components are coherent (and synchronous) with each other, while others argue that they are only loosely coupled (section 3.2.4). It is evident that there is scope for further work within this line of research as no technique available can successfully expose the nature of all components simultaneously. Each type of assessment, therefore, remains independently and equally significant; as the literature suggests [246, 7], there is still no 'gold standard' measurement of emotion. For this thesis, we focus on emotion that can be observed, irrespective of experiential (and non-visible physiological) congruence. In fact, for natural spontaneous speech, lack of coherence between measurable components may be expected. In cases of deception or social compliance, for example, the observed behaviour may not correspond to the subjective experience of an emotion. In other words, the internal state of the speaker (cause-type) may not correspond directly with the effect the characteristics of speech would be likely to have on a typical listener (effect-type) (see section 3.2.5). Moreover, we pointed out in section 3.2.6 that there are two approaches to labelling, employing a group of expert judges or working with a large-group of non-expert judges. While in most cases expert judges are assigned, the assignment of gathering large numbers of annotators, it seems, is rarely a principal research objective. Our aim is to provide effect-type labels—representing the listener's perceived expression of emotion—evaluated by an undefined group of individuals (naïve judges) through large-scale listening tasks. To this end, this brings us to our first research question:

RQ1: What are the practical prerequisites for carrying out large-scale listening tests?

Subsequently, this chapter gave an overview of discrete and dimensional type representations of emotion, including key instruments that have been used in recent studies of emotion and speech (section 3.3). First, we looked at the discrete categories approach (section 3.3.1). Language provides for a meaningful way for people to describe emotion, and tools that use single-word la-

bels remain popular. Such tools are relatively easy to develop and may be an exemplary choice when considering laymen's conceptions, given that everyday language terms are the most familiar way to describe emotions. However, the literature indicates that describing emotion scientifically with discrete labels poses several problems. Choosing an appropriate list of emotion terms appears to be the most obvious difficulty. There is a discrepancy among commentators with the type and number of categories that are compiled. This seems to be particularly complex when dealing with spontaneous naturalistic data—perhaps not so much when dealing with acted data containing stereotypical expressions. Studies that do use single-word labels often provide standardised lists that are theoretically derived or empirically established. In cases where a small amount of categories are used, such as the "Big Six emotions, it is common with naturalistic data that there are no observable occurrences in the given dataset. If one just views emotions as basic, it seems inevitable that certain underlying emotions, or emotions of the same family, will be neglected. This approach seems to inhibit progress for studies of natural spontaneous speech. Moreover, choosing too few categories can pose other shortcomings. If the task uses only a small set of categories, it is thought that the study may act as a *discrimination* task (choosing between alternatives) rather than a *recognition* task (explicitly identifying the particular category) [12, 189, 247]. In other words, if the study is a forced-choice task with a relatively small number of response alternatives, the chance of guessing correctly increases. Statistically, corrections can be made to take this into consideration. The Kappa statistic [248], for example, measures agreement on nominal data (i.e. discrete labels) and takes into account agreement occurring by chance [152, 249]. On the other hand, to avoid guessing between alternatives, some studies use free-response formats, whereby a participant can respond with a category that is not provided in the list ([250]). Studies regarding spontaneous speech often choose standard lists of non-basic emotions. This often becomes problematic as lists can easily turn out to be large, making labelling practically intractable. Moreover, ad hoc categories are often selected suited for a particular speech dataset. This, however, makes it difficult for comparing results across studies with different speech datasets, and across studies that involve various cultures. In demonstrating qualitative differentiation in acoustic patterns of emotion in speech, Scherer [12] writes that it can be problematic when emotion families are not taken

into consideration. Variants of the same emotion are generally not specified with discrete representations. On the other hand, a model that embeds emotion families within some hierarchal structure may be a suitable solution for observing variance among emotions. However, we argue that describing emotions along dimensions, the alternative descriptive framework, may be better suited for this.

In section 3.3.2, we examined the abstract and appraisal dimensional representations. First, we observed that there is a large body of empirical evidence to suggest that emotions can be successfully mapped onto *abstract* dimensions. A familiar criticism that dimensional models receive is the inability to differentiate emotions sufficiently. For this reason, common debates reflect on the type and number of dimensions needed. There are several benefits to the dimensional model. Firstly, they allow for quantitative measurements that can be tracked over time, and be visually presented, which one would argue is more beneficial for gradual transitions in dynamic stimuli [152]—although temporal measures, such as tracing, have been performed on the intensity of categorical descriptions [194]. From the literature, it is evident that discrete categories, such as the 'basic' types, are used more often in studies of facial behaviour than voice behaviour. In part, this may be due to the potential use of static stimuli (e.g. photographs), rather than merely dynamic stimuli (e.g. movies). For speech, however, only dynamic stimuli can be used, and, therefore, the development of possible temporal assessment needs to be considered. Second, dimensions indicate proximity (similarity, dissimilarity, etc.) between observed states, whereas the relationships between categorical descriptions are undefined and only identical matches are considered. As a result, dimensions give a better estimate for real consensus among raters compared with categorical descriptions [201]. Third, with a predefined list, observations will be restricted to the number in that list, even though certain other states may be observed. Dimensions, however, would not be restricted to a list and would capture all emotion types, both full-blown and underlying emotions, at least in theory. Moreover, because labels vary between studies according to the particular speech dataset, scenario, or application in mind, its subjective use poses problems for cross-corpus analyses. It has been suggested that affect dimensions are easier to match in this regard [251]. Lastly, compared to dimen-

sions, one may expect rating consistency to be higher when using emotion terms that laymen are familiar with (such as happy, sad, etc.). However, this is not necessarily the case. Some studies found that ratings on dimensions were more reliable than those for categorical descriptions [252, 221, 16]. Although many consider dimensions to be a reduced account of emotion description, many researchers aware of this limitation heedfully proceed with the dimensional approach [187, 159, 253, 157, 194, 152, 254, 32, 255, 16].

The other forms of dimensional representations examined were the appraisal dimensions (section 3.3.2). Appraisal models are increasingly becoming more accepted in the research community [151]. However, Devillers et al. [24] claim that there is a major methodological drawback with appraisal-based descriptions. That is, to annotate with appraisal dimensions reliably, the subject who is being induced needs to annotate either introspectively in real-time, which can affect the emotions experienced and expressed, or rely on the recall of the event. The distinction made between cause- and effect-type emotion descriptions (section 3.2.5) seems to be particularly worthy of attention with appraisal-based annotations. To study how emotions are generated (cause-type), these representations seem very suitable. In the case of an effect-type study, observers would be required to rate appraisal-related states in other people. So far this approach has received little empirical work [256], and remains to be examined. Such a task is, arguably, difficult to conduct successfully, which may result in low agreement levels. In fact, a study by Devillers et al. [221] compared verbal labels, abstract dimensions, and appraisal-based annotations of *perceived* audiovisual recordings, and found that agreement was relatively low for the appraisal-based annotations [221].

Embodying either the discrete or the dimensional model, in section 3.3.3 we reviewed some of the key instruments and resources used in judgement studies. Although these tools are not all made available, they do provide us with a reference point for planning and developing our own rating tool specific to our needs. In summary, we consider the following to be key aspects in choosing a descriptive framework for labelling:

Theory. A researcher needs to bear in mind the limitations that each descriptive scheme has due to theoretical issues (see section 2). If one adapts the notion of 'basic' emotions, for example, it is generally assumed that certain emotions are universal, and that they possess specific physiological and expressive profiles.

Orientation. The type of description that is required, distinguished by either cause-type or effect-type labelling (section 3.2.5), determines the type of assessment chosen for labelling emotional speech.

Accessibility. The practical usability of a measuring tool needs to be considered regarding labellers coping potential. It may be necessary to have methods that make it accessible for participants unfamiliar with emotion theory, i.e. for non-expert users.

Elicitation type. While a specific chosen set of labels may be suited for one particular speech dataset, it may not be suited for another. The labelling methods chosen often result from the type of elicitation found in a given dataset (see section 4.3 for more information). For instance, labels appropriate for acted, stereotypical expressions rarely reflect the expressions found in real-life spontaneous data.

Modality. Certain emotion terms may be modal specific (see section 4.1). Emotion terms that are particularly well suited for facial stimuli may not be relevant for speech stimuli, and vice versa. For example, studies have labelled speech material for underlying states (or attitudes) such as 'motherese' [193] (baby-talk) and 'sarcasm' [257]. These terms are arguably only observed from speech—specific to the dataset under investigation—and not from facial expressions, at least not to the same degree. Most studies of spontaneous speech consider strongly the notion of underlying emotions. Studies of facial behaviour, however, rarely consider such underlying emotions. The distinction between the different modalities of expression, and the relevant emotion terms used, is an important one to consider when choosing a labelling method. Although the discrepancy between emotion lists appear to exists between modalities studied, to our knowledge this is not widely acknowledged by others in the literature.

Lastly, we reviewed several recent schemes proposed for labelling naturalistic emotional speech datasets, and outlined some of the difficulties that researchers were faced with (section 3.3.4). It appears that labelling methods proposed for naturalistic speech are, in most cases, data-driven. Several propose comprehensive schemes that provide in-depth qualitative distinctions in conjunction with dimensional labelling (Evaluation and Activation). It is known that Evaluation and Activation dimensions capture a relatively large amount of emotion variation [201]. The use of these dimensions appear to be suitable for natural spontaneous speech, as demonstrated by the work on the Belfast Naturalistic database [244] and the JST/CREST Expressive Speech Corpus [245]. The descriptive scheme chosen for this thesis are the two dimensions Activation and Evaluation. By taking into account some of the issues outlined by Cowie et al. [30] (see section 3.3.4), we postulate this choice for the following reasons:

Economy. The divergence about the number of emotion words used and the divergence about their primitive significance makes the discrete representation inconvenient.

Consistency. The different emotion categories proposed relevant to a particular speech dataset, using *ad hoc* or *cover classes*, limit the comparison of results across studies. The objective nature of dimensions is, arguably, more advantageous for this reason.

Intermediates. Naturalistic speech databases often contain subtle (intermediate or combined) states for which there is no appropriate single-word term. The broad coverage of the dimensional representation would capture all potential intermediate emotions, albeit a lower degree of qualitative differentiation.

Furthermore, although discrete representations are more suitable for qualitative differentiation, we consider that one of the major advantages of using abstract dimensions is:

Quantifiability. The quantitative nature of dimensions provide coordinates for proximity measurements for more accurate rater consensus estimates, and provides coordinates that can be visually presented and tracked over time. A dimensional representation is more sufficient for analysing temporal/dynamic changes.

However, there are two things we need to consider for the task of labelling. First, although naturalistic emotional speech has been adequately evaluated on dimensions by expert judges familiar with the concept of dimensional theories, it is uncertain as to how appropriate this method is for naïve judges. In order to appoint naïve listeners, we need to consider the accessibility of this concept, how well do lay people understand the concept of emotional Activation and Evaluation. Second, because the speech material is of a naturalistic sort, emotions are mostly underlying and not as clear-cut as might be expected, possibly making the task too demanding for naïve judges. Although the Activation and Evaluation dimensions have been used successfully with some datasets, this may not be the case for speech composed of mood inducing procedures. With this in mind, this gives rise to the following research question:

> **RQ2:** Can listeners adequately capture variation of Activation and Evaluation of emotion in naturalistic speech?

3.5 Conclusion

This review chapter considered two essential requirements involved with labelling emotional data, (1) how to measure and assess it, and (2) how to conceptualise and classify what has been measured in order to make emotional states distinguishable. First, the different assessment techniques (self-reports, physiological measurements, and behavioural observations) were examined, outlining that each technique deals with a particular component of emotion: cognitive, experiential, physiological, or behavioural (section 3.2). Research in this thesis is concerned with labelling of *effect-type* description (section 3.2.5), whereby behavioural measurements (of perceived expressions) are exclusive. Additionally, it was argued that labels provided by non-expert judges are equally valid. We have decided to carry out a case study (Chapter 7) in order to perform large-scale listening tests, thus exploring research question one (RQ1).

Second, this chapter considered the different prevalent approaches for classifying emotions (section 3.3), which can be broadly distinguished by discrete or dimensional theory. We con-

sider the Activation and Evaluation dimensions to be the most enduring in the literature, and appropriate for dealing with naturalistic emotional speech. To implement this into a tool, the case study will examine if viable effect-type labels can be determined. This task will, in and of itself, examine whether the use of dimensions is suitable for naïve judges, and investigate if a sufficient amount of emotion is present in terms of these dimensions in the given naturalistic emotional speech dataset, composed of mood inducing procedures. This work will contribute to attempting to answer research questions two and three (RQ2 and RQ3).

Although throughout this chapter we have emphasised that the issues with data labelling interconnect deeply with the issues of speech data types, little has been mentioned about the available speech data itself. This will be reviewed in the next chapter. Furthermore, to give our investigation of data labelling more significance, the following chapters will link in with our proposed objectives, making the connection with the perception process of vocal expression (see Chapter 4), and the acoustic correlates of vocal expression (see Chapter 5).

4

Emotion and Speech

Having looked at a broad overview of the theoretical foundations of emotion in Chapter 2, and how we may describe emotion in Chapter 3, this chapter deals with how emotion is communicated through speech. Speech is an acoustically rich signal that comprises several layers of information: linguistic, paralinguistic, and extralinguistic [258]. The linguistic aspect communicates the verbal coding system of human language, the paralinguistic aspect communicates non-verbal information about the speaker's feelings, attitude, or emotional state, and the extralinguistic layer conveys information about a speaker's characteristics, such as identity and gender. These three communicative functions are an integral part of speech that are characterised by certain acoustical patterns, yet are intertwined within the same speech signal. First, this chapter will give an overview of the different research studies that one can undertake with regard to vocal communication of emotion, with emphasis on labelling paralinguistic mean-

ing. Second, in order to conduct research on emotion in speech an obvious starting point is the acquisition of emotional speech data. For this, we will summarise the different types of data available for emotion in speech research, and the main issues that need to be considered.

4.1 Expression and Perception

When humans interact, emotion is communicated innately through various modalities (i.e. facial expression, speech, and body gestures). Recently, an increasing number of studies report on multimodal recognition of affect and the existence of relations and/or dependencies between the different modalities [23, 259, 260]. A study by Busso and Narayanan [226], for example, showed that if one modality is constrained, other channels have a stronger emotional modulation. Furthermore, there appears to be attentional biases towards facial expressions and vocal expressions when inferring emotion. Some suggest that the visual channel is the most common way for people to infer emotions in everyday life [261, 262], while others emphasise vocal cues are [263]. There is an increasing amount of research being carried out on how emotion is conveyed and perceived from vocal emotion expression, and it has been shown that listeners are quite successful at inferring affective states and attitudes on the basis of vocal cues alone [264, 265, 266, 73, 267, 268], with an accuracy percentage generally found at around 50% [189]. For acted vocal portrayals, Scherer [12] reported in his review recognition accuracies between 55% and 65%—being five to six times higher than expected by chance—while the general reported average accuracy for facial expressions is around 75%. He points out several potential reasons for this difference. First, the dynamic nature of vocal stimuli may produce more complex patterns that are less distinguishable compared to specific muscle configurations found in static facial stimuli. Second, emotions that are from a similar family may be more distinctly recognised when expressed vocally. He further pointed out that emotions such as sadness, anger, and fear are generally recognised more accurately from vocal expressions, yet joy seems more ambiguous for vocal expressions compared with facial expressions, from which it is recognised almost perfectly. In fact, specific emotions may be better expressed via the voice channel due to the situational context. For instance, if distance governs emotional

communication, one is more capable of expressing emotions such as 'fear' vocally—serving as an alarm signal [269]. Nevertheless, it is widely acknowledged that the human voice is one of the primary channels of social and affective communication [105]. Darwin [33] provided the first comprehensive account of vocal emotion expression. His perspective, consistent with other contemporary perspectives (e.g. [73, 189]), suggests that specific vocal acoustic cues are associated with discrete affective states (see [269] and [147] for a review on acoustic correlates for discrete emotional states). In more recent studies, vocal acoustic cues have also been shown to correlate with a small set of continuous dimensions [157, 270, 87].

4.1.1 Brunswikian Lens Model

Vocal communication of emotion typically involves two or more individuals emitting and registering signs of emotion. The various communicative aspects can be illustrated within the *Brunswikian lens model*, a conceptual framework of perception [271]. The model was originally intended for visual perception, which has since been used to conduct judgement and interpersonal perception studies [272, 273]. It is based on *cue theory*, which suggests that emotion is perceived through unconscious inferences drawn from a combination of sensory cues that are *probabilistic* and not fully reliable [242, 274]. Brunswik suggested that objects are frequently observed from multiple cues, and emphasises that the observer makes do with a variable set of constricted and/or imperfect cues. In other words, the perceiver can intuitively make valid interpretations from uncertain information conveyed by these cues. The perceiver's flexible ability to substitute interchangeable cues is referred to as *vicarious functioning*. The matter of investigating individual cues is complicated by the fact that configurations associated with one particular emotion might, in fact, be associated with multiple other behavioural states, which are not necessarily emotional ones. Particular cues may have alternative meanings. For instance, a smile can be associated with happiness and success, yet be similarly associated with emotions of hedonic contrast, such as nervousness and failure [138]. This issue has been referred to as *systematic ambiguity* [19, 156].

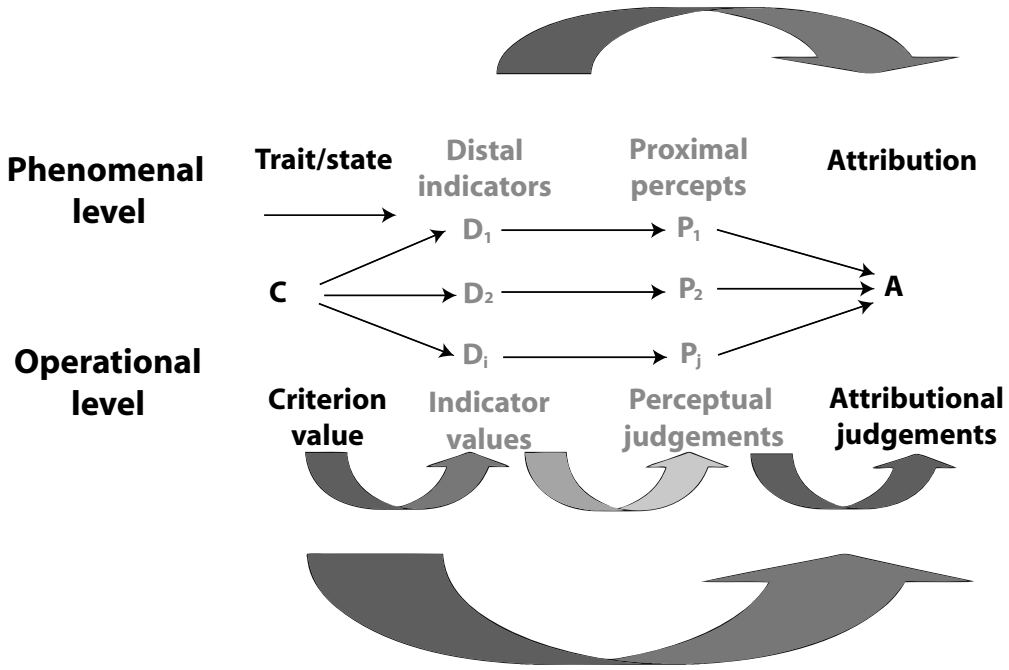


Figure 4.1: Scherer's modified version of the Brunswik model of perception [12]. This figure shows the relationships between objects and processes involved on the "phenomenal level" (top) and the corresponding "operational level" (bottom).

Based on Brunswik's functional lens model, Scherer [73] suggested that research should be based on a modified version of this model that tailors vocal communication of emotion (Fig 4.1). In a more recent review of the different paradigms of emotion and speech, he has re-asserted this view [12]. The model illustrates the different aspects of emotion communication, which distinguishes between the speaker's *encoding* (or expression) of emotion, the *transmission* of the sound, and the *decoding* (impression) performed by the listener. According to the model, the perception process involves *distal* and *proximal* cues. A proximal cue (cues close to the observer) is a physical stimulation pattern in the organism's senses that corresponds to the cues of any given external object or event in the environment, i.e. the distal cues (cues distant from the observer) [275]. In terms of the visual domain, distal variables include measurements of an object's size, distance and position, whereas the proximal variables are represented by the retinal image it produces. The example provided by Scherer [12] for auditory perception explains that distal cues are the objectively measurable acoustic changes (cues distant from the observer), such as the fundamental frequency of speech. These cues, in turn, are transmitted

through a medium to the listener, which are then represented as proximal cues. The proximal cues can be explained by the process that "gives rise to the pattern of vibration along the basilar membrane, and in turn, the pattern of excitation along the inner hair cells, the consequent excitation of the auditory neurons, and, finally, its representation in the auditory cortex". In consequence, the listener attributes the perceived proximal cues to the speaker's emotional state.

In summary, the model highlights the uncertain relationship between objectively measured (distal) and perceived (proximal) cues. In real life, the distal cues are often misrepresented due to the distortion of the acoustic signal through the transmission channel (affected by noise, distance, or medium), and/or by the individual's physical characteristics of their auditory perceptual system (e.g. attenuation of certain frequency bands due to hearing impairments). These variables affect the transduction and coding process. Moreover, even if the proximal cues reliably map the valid distal cues, it is still possible for the listener to infer incorrectly the speaker's emotional state due to cognitive influences during encoding. To quantify the relationships between distal and proximal variables, the relationship between acoustic patterns and the underlying speaker state, Brunswik suggested that a *correlation coefficient* is the most appropriate measurement, the degree of which offers an index for the *ecological validity*[1] [274]. Most studies combine the detail of encoding and decoding processes, while others focus specifically on one aspect of the perceptual inference process. Scherer distinguishes the type of studies that fit the various aspects outlined within the model. These are outlined below.

4.1.2 Encoding Studies

In encoding studies (referred to as 'speaker-centred' studies by Schröder [87]), the research is exclusively focused on establishing the association between emotional states and the measurable acoustic parameters of the speech signal—assuming that unique acoustical patterns (distal cues) exist for the different emotional states. Studies regarding automatic recognition of emo-

[1] Ecological validity is the extent to which the conditions simulated in an experiment or study reflect the conditions in real life. In other words, it is the extent to which the findings can be generalised to the real world.

tion in speech, for example, are considered to be placed in this domain [87]. Because voice analysis is representative of the type of speech data under investigation, encoding studies can be grouped into three major categories: acted emotional expression, natural emotional expression, or induced emotional expression [12]. That is, the different data types (as presented in section 4.3) are characterised according to the emotion elicitation found in the respective data types. That being the case, studies according to these data types are methodologically different. First, they can be distinguished by the degree of experimental control throughout the recording process. Whereas recorded material found in naturally occurring situations is mostly beyond the researcher's control, induced and acted data types are generally obtained in a controlled environment. As a result, it is rather easy to obtain good audio quality on the recordings of acted and induced data types. Second, both natural and induced data types are oriented towards ecological validity, representative of actual emotional occurrences, while acted data represent either stereotypical expressions or try to *reproduce* actual real-life expressions. Essentially, encoding studies involve finding associations between a given emotion label with a set of acoustic parameters. For acted data, the instruction is given to portray a given emotion, so the intended emotion is predetermined. Therefore, no perception tests need to be carried out to determine the representative labels—although perception tests can be performed to determine the quality of acting. With natural and induced data, on the other hand, there is no certainty with the precise nature of the underlying emotion, so perception tests are necessary. In other words, most studies involving natural and induced data types combine both encoding and decoding aspects in one study. Third, the issues associated with labelling often rests on the type of data gathered. In recent studies that availed of acted material, only a few 'basic' emotions were studied, but, as previously mentioned in section 3.3.1, for natural spontaneous speech data it is common to find an approach to labelling that is data-driven.

Essentially, the task for emotion recognition involves training models to perform automatic classification of emotions. In order to do so, a training set of labelled speech is needed. It should, therefore, perhaps be mentioned that the performance of an automatic classifier relies not only on the characteristics of the given speech data, but also on the precision in the labelling

of states. Therefore, one must aim to choose labels that are relevant and appropriate for a given speech dataset, and that the labels provided are trustworthy, the classic criteria being validity and reliability [188]. Likewise, when assessing the trustworthiness of a label, one should refer to the distinction between cause- and effect-type labels—the distinction which conceptualises that a listener's perception of emotion does not necessarily imply attribution of a speaker's experienced emotion (see section 3.2.5). Cowie et al. [240] explain that for cause-type labelling, for example, it is appropriate to seek the 'ground truth', which establishes what the speaker's state was at the time of speaking, while for effect-type labelling it is critical to have the degree of rater agreement associated with the label.

4.1.3 Decoding Studies

Decoding studies examine the listener's ability to recognise the emotional state from a speaker's vocal expression independent of (normally constitutive) lexical content. Decoding studies are an essential starting point in the study of vocal expression. Prior to investigating vocal correlates of emotion, decoding studies establish if a speech sample actually conveys emotion that can be reliably recognised by listeners. In most previous decoding studies, listener judgements were examined on acted speech. Speech samples with acted emotion can be produced in such a way that linguistic information does not influence the judgement of emotional meaning (e.g. using meaningless sentences), and thus the label is independently derived from paralinguistic patterns (vocal expression that refers to qualities of speech rather than verbal content). The actors are generally instructed to portray a predefined set of emotions, so the intended labels are assigned beforehand. Subsequent listening tests can then demonstrate if listeners can accurately perceive the intended emotion portrayed. One of the major drawbacks with decoding studies is that the expressed emotions are limited to the ones instructed to act, which is quite different to the complexity and number of emotions found in natural speech. In addition, the number of response alternatives typically provided is often too few (if only 4-6 response alternatives, for example). That being the case, the findings are then characterised as *discrimination* performances (using exclusion and probability rules to guess the right answer) rather than *recognition* per-

formances [189], which would compromise ecological validity. It is questionable, therefore, if such observations can be generalised to real-life scenarios. The chance of guessing the answer correctly clearly depends on the response options provided [276, p. 257], however, researchers can correct the accuracy coefficients to compensate for this effect [277, 12, 65]. In Scherer's review [12], recognition rates for facial expressions yielded average accuracy results of around 75%, while recognition accuracies for vocal expression were reported to be between 56% and 65% (around five times higher than would be expected if randomly guessed). It is explained that these differences could be due to several factors. For example, the dynamic nature of vocal stimuli is less inclined to produce stable acoustic patterns compared to basic facial muscle configurations (using static stimulus material like photos), and members of a similar emotion family can be more distinct from each other vocally in comparison to facial expressions.

The recognition accuracies for vocal expressions appear to be comparable in studies conducted cross-culturally, indicating the universality in emotion inference. Across nine different countries in Europe, the United States, and Asia, Scherer [43] showed that the overall recognition rates for vocal emotion portrayals of four different emotions, and neutral state, is 66%—the stimuli used were meaningless multi-language sentences. The claim that emotions can be recognised with better than chance accuracy has been supported by a more recent meta-analysis by Juslin and Laukka, including within- and cross-cultural studies [242][2]. Out of 104 studies reviewed, the summarised data included 1 to 15 emotions, 87% of the studies included acted expressions, 13% used manipulated speech (see *cue masking*, section 4.1.4), 7% used mood inducing procedures, and 12% used natural spontaneous speech. Whereas previous reports generally reflect emotional portrayals, they reported that decoding accuracy rates for natural expressions were similarly well above chance. However, they acknowledged that the number of studies using natural speech was too few, and thus no definitive conclusions could be made. Furthermore, accuracy estimates across the different studies were similar for individual emotions. Sadness and anger were generally better perceived than fear, happiness, and tenderness.

[2]The data reviewed was summarised in terms of Rosenthal and Rubin's [278] effect size index for one-sample, multiple response alternatives, pi (π). Regardless of the number of response alternatives, the index allows accuracy scores to be transformed to a standard scale (.50 is the null value, and 1.00 represents 100% correct)

4.1.4 Inference Studies

Through listening tests, decoding studies and inference studies both focus on the listener's judgement of speech. While decoding studies establish the listener's ability to recognise an emotional state from a speech sample (decoding), inference studies focus on the specific voice cues (distal) that a listener utilises in the emotion inference process. For this purpose, listener's ratings are correlated with voice cues that have been measured or manipulated. This area consists of three major methods: (1) cue measurement and regression (2) cue masking, and (3) cue manipulation via synthesis.

Cue Measurement and Statistical Association

Cue measurement and statistical association involve correlating the listener's judgements of the speaker's emotional state with the measured acoustic characteristics (e.g. [31, 279, 280, 281]). A study by Banse and Scherer combined both encoding and decoding aspects in a study, where they used 14 acted portrayals of emotions (encoding), and subsequently presented the recorded material to judges for rating (decoding). This allowed them to determine recognition accuracies of the target emotions, and specify the effect of different acoustic cues on emotion inferences based on both their correct and incorrect responses. They used multiple regression analysis to investigate the relation of a judge's emotion inferences with the acoustic measurements. They found that the variance in judgement could be explained by 9-10 acoustic measurements, which included F_0 mean, F_0 standard deviation, mean energy, duration of voiced periods, the proportion of voiced energy up to 1000Hz, and spectral slope up to 100Hz [12].

These results correspond to judgements made on acted portrayals. With natural spontaneous data, associating the acoustic cues with the listener's judgement of a speech sample would be the preferred method, as the target emotion is generally not known. Likewise, multiple regression analysis can be conducted on listener ratings based on dimensions [157], or on discrete categories of relatively mild affective states such as irritation and resignation [31]. The

study by Mori et al. [157] found that peak intensity and F_0 range were highly correlated with most dimensions used (pleasant–unpleasant; aroused–sleepy; dominant–submissive; credible–doubtful; interested–indifferent; positive–negative), and a voice quality parameter correlated for some dimensions but was dependent on the speaker. Laukka et al. [31] included features related to pitch, intensity, formants, voice source, and temporal aspects of speech, and found that listener ratings of irritation, resignation, neutral, and emotion intensities were associated with them.

Cue Manipulation via Synthesis

Cue manipulation via synthesis has allowed researchers to perform perception tests on stimuli where various acoustic parameters can be controlled within a speech synthesis system (e.g. [87, 282, 283]). Broadly speaking, there are three different techniques to create synthesised speech, each with limitations and advantages depending on the researcher's aim (see [247] for a more detailed review):

- *Formant synthesis (rule-based synthesis)* generates speech based on rules that simulate acoustic properties of human speech production. Formant synthesis does not process human speech samples at run time, which makes it computationally favourable as it requires little memory. In addition, this type of synthesis gives a high degree of control over parameters related to voice source and vocal tract, which makes it particularly suitable for those investigating specific cues that a listener may utilise to infer emotion. However, such systems, so far, generate relatively unnatural speech that sound robot-like compared to concatenative systems.

- *Diphone concatenation* uses a series of human speech recordings from a database to generate a synthesised output—the database holds diphones, which are stretches of speech from the middle of one phone (speech sound) to the middle of the next one. The synthesis produces the required F_0 contours through signal processing techniques, which can create some distortion. This technique is normally considered more natural than formant synthesis. However, because it uses speech samples within databases, it is computation-

ally less resourceful. Furthermore, most diphone systems only allow for control over F_0, duration, and intensity. In most cases, there is no control over voice quality.

- *Unit selection* is a corpus-based technique that uses relatively large amounts (i.e. several hours) of recorded speech (or speech units) from a database. In some cases where well-matching units are found, no signal processing is necessary. In cases where there is signal processing, the parameters defined can be the same as for diphone synthesis. This technique is generally perceived as the most natural.

Speech synthesis is a useful tool for studying emotion and speech. It has been successfully utilised to specify which vocal cues (distal indicators) are used in emotion inferences [284, 285, 286, 287]. The premises for emotional expression used in these studies are generally acquired from literature reviews and/or by trial-and-error adjustments based on listening tests [29].

Cue Masking

Cue masking (also referred to as content-masking) procedures modify and/or remove vocal cues from the speech signal. They have mostly been used to mask verbal content (i.e. to render speech unintelligible), which can then be used to investigate vocal expressions independent of verbal information. The main advantage to this method is that intelligibility can be removed from all speech data types (acted, induced, or natural). Therefore, these techniques seem particularly useful for judgement and analysis studies of natural spontaneous speech [288], for which the verbal content cannot be controlled by means used for producing acted material. Inevitably, applying these methods will affect certain acoustic properties that may characterise the emotion present. So, in practice, it is impossible to preserve all acoustic information while at the same time remove all verbal information [145]. However, the act of modifying, masking, or isolating specific cues can, in its own right, be used to investigate systematically how different voice cues contribute to the emotion inference process [29]. Examples of procedures include:

- *Low-pass filtering* [146, 145] removes the higher frequencies above a particular cutoff point, which are important for speech intelligibility, while letting the lower frequencies

of speech to pass through. Low-pass filtering preserves tonal (e.g. F_0 contours) and sequential properties (e.g. speech rate), yet it reduces spectral content (crucial for voice quality) and attenuates perceived loudness.

- *Random splicing* [289, 290] divides speech samples into smaller segments and randomly rearranges them. This method, therefore, disrupts the temporal and sequential organisation (e.g. pausing, F_0 contour continuity), but it retains spectral content important for voice quality information (e.g. harshness, denasalised).

- *Backward speech* [291, 292] can be achieved from playing tapes backwards or, nowadays, simply digitally reversed. It seems that some phonetic and indexical information is retained. Although it generates reversed intonation contours [293], features such as mean pitch, pitch range, and some aspects of vowel and voice quality are preserved [291].

- *Pitch inversion* is a method used less often. The study by Scherer [294] inverted the frequencies using a balanced modulator to fold the audio spectrum around a carrier frequency. This technique degrades the normal harmonic relationships necessary for intelligibility but retains stress patterns and intensity contours.

- *Tone-silence sequences* [294] is a method used for audiotapes to generate a sequence of tones and intervening silences to replicate the original sequence of speech sounds and silences. Again, nowadays, digital signal processing techniques allow for a much simpler and more effective way of manipulating audio speech material in a similar manner (e.g. audio to midi/synth conversions).

- *Reiterant speech* [295] is produced by substituting the original syllables of an utterance with other syllable imitations with similar F_0 contour. If the goal of a study is to preserve as much of the original speech stimuli as possible, and not to isolate or modify cues, this method is probably the best as it preserves F_0, temporal, and voice quality measurements [29].

4.1.5 Transmission Studies

As previously mentioned, an important aspect of the Brunswikian lens model is that it highlights the fact that the perceptual representation of cues (proximal percepts) may not correspond in a one-to-one mapping with the objectively measured cues (distal indicators). The model explains how the transmission of the distal signals from the sender to the receiver, where they are represented on a subjective proximal level, may be responsible for inference inaccuracies and should therefore be examined separately. In real-life situations, the transmission process will, at least to some extent, degrade the acoustic signal (distal cues) due to (1) the physical space through which the sound is transmitted, and (2) the transform functions in perception, influenced by the individual's auditory perceptual system. For the former, this may include aspects such as distance between the sender and the receiver (weakening the signal), the presence of interfering environmental sounds (affecting speech production and perception effort), signalling mediums such as telephone lines (thus limiting the frequency range), or obstructing materials such as walls (attenuating certain frequencies). For the latter, the transformation of the distal signal converted by the individual's physical characteristics of their auditory perceptual system to a perception can also become misrepresented. Such modelling is based on work in the psychoacoustics of speech perception. The first example Scherer [12] gives is that of perceived loudness correlating more strongly with the amplitude of a few harmonics or even a single harmonic rather than with the overall intensity. The second example has shown that listeners are able to distinguish vocal effort produced from both loud and soft voices but presented at the same perceived loudness, indicating that listeners seem to have an internal model of a specific spectral distribution representing vocal effort. Similar effects can be shown for the perception of F_0 contours, loudness, and duration of spoken utterances.

4.1.6 Representation Studies

Scherer [12] explains that the reception of the sound signal by the auditory perceptual system is stored in short-term memory as a representation of the proximal cues, at which point attributional judgements are made. Within the framework of the Brunswikian lens model, rep-

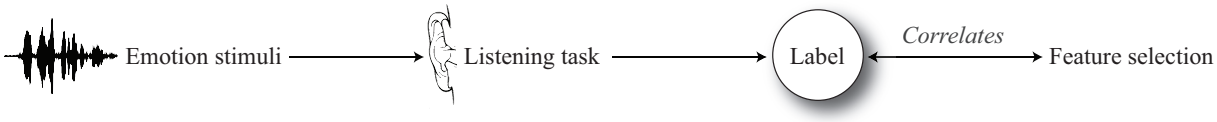


Figure 4.2: General procedures for the automatic recognition of emotion.

resentation studies try to make measurements of these proximal cues from verbal reports of a listener's subjective impression of the vocal qualities. Accordingly, the principal area lies with the inference algorithms that are made by listeners, of which vocal impressions act as indicators of particular emotion types. One of the major difficulties with representation studies is that verbal reports for voice qualities are constrained to the semantic categories available in a given language. There are few words to describe voice qualities that are used regularly, neither in everyday language nor in the literature. Example words that describe voice qualities include: nasal, sharp, and rough. Moreover, it seems difficult to obtain high inter-rater agreement measures for many of the categories used in such studies, which may be due to the insufficient use of them in normal everyday language. Incidentally, there are currently very few representation studies that study the inference structures used by listeners.

4.2 Prosody and Semantic content

When investigating emotion in speech recognition, labels are an essential requirement for researchers to explore the relationship between emotion and speech (see Chapter 3 for more information). The labels that index the expressed emotion are, ultimately, associated with a certain facet drawn from the speech signal. In many cases, labels are determined from the ratings obtained from listening tasks. The perception of emotion is based on prosody and semantic content, however. In this section, we provide a distinction between possible approaches regarding the labelling process from judgment tasks.

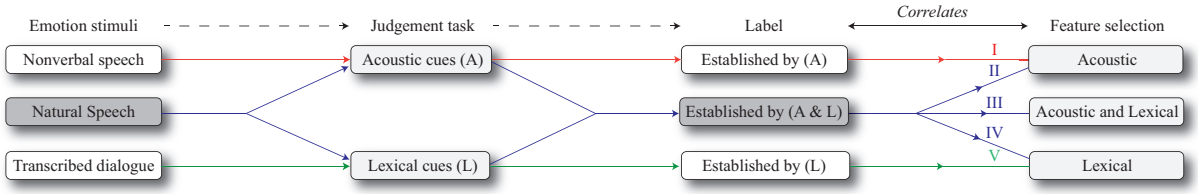


Figure 4.3: A distinction between possible research orientations with reference to channel contributions in emotion communication.

4.2.1 Labelling Precision

To build a model to automatically recognise emotion from a speech signal, an association needs to be determined by (1) the selected features and (2) the label that represents the expressed emotion. The procedure is briefly illustrated in Figure 4.2. First, features are extracted and analysed to determine the existence of any meaningful patterns. Much work focuses solely on correlating acoustical features with an expressed emotion [296, 297, 298, 280, 299, 300, 301, 302, 303, 304, 305, 306]. It is well known, after all, that certain acoustic parameters correlate with emotion. However, we can gather information about the acoustic and linguistic aspects of a speech signal. Some approaches emphasise the influence of linguistic (or textual) content by itself [307, 308, 309, 310], whereas other efforts select and augment both acoustic and linguistic features [193, 311, 312, 313, 314, 315]. Incidentally, there are studies that incorporate discourse context (dialogue acts) as a third source of information [316, 249]. Most of these studies that investigate combined knowledge sources report on an improvement in automatic emotion recognition rates. Second, a predetermined label specifies the expressed emotion that the extracted features are correlated with. Generally, the label is derived by humans performing subjective listening tests (as performed in Chapter 7). Often labels are derived from speech that comprises contributing acoustic and linguistic cues—listeners integrate both sources to infer the expressed emotion. At this stage, one may investigate a relationship between the label and the acoustic (or linguistic) channel. However, the labels are often not derived solely from acoustic (or linguistic) cues.

Craggs [310] argues in the context of linguistic annotation that if the provided labels are not based exclusively on the information found in the linguistic content, one should question the

reliability on the conclusions drawn about a relationship with those labels. Likewise, one could say the same when determining acoustic correlates. With this in mind, we expand on the concept illustrated in Figure 4.2 to show several possible research orientations that correlate extracted features with an expressed emotion (see Figure 4.3). The diagram illustrates that labels are derived from a set of features based on acoustic or linguistic cues. For example, study type I in Figure 4.3 illustrates that acoustic features are correlated with a label based solely on the information of the acoustic cues, i.e. from nonverbal speech.

As mentioned above, substantial work has been carried out on natural speech that investigated acoustic (see II in Figure 4.3), lexical (see IV in Figure 4.3), or a combination thereof (see III in Figure 4.3) to determine an optimal set of correlated features with a predetermined label. However, as the diagram shows, the labels are derived from both acoustic and linguistic information. As illustrated, the difficulty with natural spontaneous speech is that labels are generally established from both acoustic and linguistic cues. Clearly, each channel has a specific contribution in affective speech processing, but the level of significance of each and to which each channel aligns and interacts with the other remains to be investigated. One channel can serve to elaborate on or conflict (congruence or incongruence) with the conveyed meaning of the other channel, such as sarcasm [257] or in cases of deception (see section 4.3.4). The study of channel alignment is not straightforward and there are unavoidable trade-offs that a researcher has to make when mitigating or controlling one aspect of speech—although, these constraints are less severe with non-spontaneous speech.

To study the linguistic content independently, one can measure the effect transcribed text has on the inference of emotion. Because the information is in that case delivered in isolation, the derived label is based solely on the linguistic content (see V in Figure 4.3). It would, therefore, be more justifiable to draw conclusions about the relationship between the linguistic features and the derived label. However, transcribed text may not have the same influence as spoken text, as it may not have the same ecological relevance [317]. To address this, studies have presented spoken words, either in a tone congruent or incongruent to the words, to measure

their effect on a listener's subjective ratings [318, 319, 317, 320]. On the one hand, it has been shown that lexical content of spoken words plays a more significant role in conveying emotional information over tone of voice [321], while on the other hand, it has been suggested that prosodic cues are mostly used to identify with an expressed emotion [322]—although these studies might be culture specific. Methodologies that control (or manipulate) semantic and/or acoustic content are generally restricted to acted speech, but have also been performed using synthesised speech, therefore, impracticable for spontaneous speech. A mixed control study conducted by Ilves [285] used synthesised speech to control for both prosodic and semantic cues. This study investigated how synthetic verbal stimulation with emotional content would affect the perceiver. The findings showed the ratings of emotions were significantly affected by the emotional content of the synthesised words. Moreover, the study also showed that the quality of the same voice received higher ratings when the verbal content was positive in comparison to the neutral and negative sentences.

4.2.2 Masking Linguistic Content

For speech that is truly natural and spontaneous, it is impracticable to control the linguistic content by scripting its spoken dialogue. To separate the acoustic and linguistic channels, researchers can, instead, render speech incomprehensible using masking techniques such as random splicing [289], backward speech [291], pitch inversion [294], and foreign speech [323], and re-entrant (REENT) speech [324] (study I in Figure 4.3). Inevitably, using these methods to mask linguistic content will affect certain acoustic properties that may characterise the emotion present, e.g. backward speech will have reversed intonation contours. It should perhaps be noted that, nowadays, we benefit from technological advancements compared to early research that availed of tape to manipulate speech and was often subject to errors by simultaneous pitch changes.

Another widely used method for masking linguistic content, adapted from research into speech intelligibility, is low-pass filtering [146, 325]. Low-pass filtering removes high frequencies,

which are important for speech comprehension (i.e. intelligibility), while leaving the lower frequency regions of speech intact. Low-pass filtering preserves tonal and sequential speech variables (or prosodic characteristics) such as speech rate, rhythm, intonation contours, and stress patterns [326, 293]. Tonal aspects of speech, which are dominated by the lower frequencies, play an important—if not the most important—role in affective expression [326]. An emotional judgement study that uses low-pass filtered speech can remove linguistic content and assess the role of the remaining vocal parameters that listeners base their judgement of emotion on.

The precise nature of low-pass filtering and its impact on speech perception remains to be established in studies of speech intelligibility, and emotion in speech. To our knowledge, few recent studies have used low-pass filtering as a masking technique on 'spontaneous' speech. One such study by McNally et al. [325] elicited emotion in participants—patients with panic disorder, major depressive disorder, social phobia, and healthy control participants—by asking them to recall both fear and neutral autobiographical memories. The speech clips were recorded onto audiotape, rather than being digitised, and content-filtered to eliminate frequencies above 400Hz. Each clip was evaluated by raters along the widely used scales: *negative*, *aroused*, and *dominant*. Two added scales, *anxious* and *sad*, were chosen applicable to the type of speech material being rated, i.e. speech recordings from patients with mood and anxiety disorders. For the dimensions they studied, content-filtered speech conveyed enough information on fear related emotional valence. Similarly, Knoll et al. [146] studied perceptual ratings of vocal effect on filtered speech directed at Infants (IDS), Adults (ADS) and Foreigners (FDS). Raters were questioned on four scales: *positive* vocal affect, *negative* vocal effect, *encouragement of attention*, and *comforting and soothing*. The authors noted that certain affective scales might be more informative for a particular type of speech. Thus, comforting and soothing, for example, might be more relevant for Infant Directed Speech (IDS). Four different filter conditions were investigated. It was acknowledged that cutoff frequencies above 1000Hz kept some semantic information discernible that may have confounding effects on the rater's perception.

The effect of the low-pass filter on intelligibility mainly depends on the selection of the cutoff frequency. The range that is important for speech intelligibility is about 500 to 5000Hz [327]. However, there does not seem to be a standard optimal cutoff point for intelligibility as this may vary depending on the proportion of background noise, quality of recording, and the speech characteristics of the speaker. Research investigating speech intelligibility filter speech into a number of frequency bands called 'analysis filters' [328, 329]. In most cases of nonverbal communication, however, researchers use a *single* fixed value ranging from 300Hz-600Hz [330, 325, 144, 145, 293], or a *set* of fixed cutoff values in order to compare the effect of each condition e.g. [146]. MacCallum et al. [331] recommend the cutoff frequency to be at least one octave above the F_0 (minimum of 300Hz) to ensure acoustical analysis accuracy of percent jitter, percent shimmer, fundamental frequency (F_0), signal-to-noise ratio (SNR), and nonlinear dynamic measures (correlation dimension and second-order entropy).

4.3 Emotional Speech Acquisition

The most important initial step in building an emotional speech corpus is, of course, collecting appropriate speech material that can be labelled for emotion. Obtaining high-quality emotional speech data is not only a technical problem, but also a practical and theoretical one. Currently, available speech materials are inevitably diverse in nature, which makes it particularly difficult for any cross-corpus analysis [332]. The type of speech material acquired in a given study is distinctly representative of its goal. When collecting speech material, central to the experimental design are considerations of audio quality and emotion eliciting processes. The audio quality of recorded speech material needs to be sufficiently adequate for acoustic analysis as certain voice measures are error prone to poor audio quality [29]. Conditional on recording setups and environments, technical issues, such as background noise, reverb/echo, and microphone placement ultimately affect the desired signal. Furthermore, the circumstances that the elicited emotion is characterised by is a topic that has been widely discussed [333], the problem of authenticity being central to the matter. Picard et al [334] give a summary of five conditions that characterise the elicited emotion, which influences the acquisition of speech data:

- *Subject-elicited* vs. *event-elicited*: is the elicited emotion deliberate (i.e. is the emotion posed) or is the elicited emotion an outcome of a stimulus or situation outside of the subject's control (i.e. is the emotion spontaneous)?
- *Lab setting* vs. *real-world*: is the recording environment in a lab or in a setting natural to the subject?
- *Expression* vs. *feeling*: is the emphasis of a study on external expression or on internal feeling?
- *Open-recording* vs. *hidden-recording*: is the subject aware of being recorded?
- *Emotion-purpose* vs. *other-purpose*: is the subject aware that the experiment is about emotion?

The type of envisioned application determines the relevance of the above factors. If the application is for acted emotion in movies and animations, for example, one is generally concerned with the quality of acting, and most of the factors above may not be of immediate concern. For real-life applications, all factors are generally considered during the experimental design. Whether one is interested in the external expression (how emotion is observed) or internal feeling (the ground truth)[3], for real-life applications, methodologies tend to be oriented towards ecological validity. In an ecologically valid setting, the ideal situation is to capture actual emotional occurrence in a subject in response to a personally significant circumstance in a real-world situation where the recording equipment is hidden and the subject is unaware of the experimental objectives. As one can imagine, such situations are difficult to improvise because of ethical reasons. With these specifications in mind, speech data can be broadly grouped into three types of expressions: simulated, natural, and induced. It is important to be aware of the distinction between these types, and their potential contributions and limitations. Several extensive reviews of available emotional speech corpora have been provided [240, 241, 242, 201]. The reviews show that the majority of corpora so far consisted of acted representations of emotion, indicating that the majority of results of previous research do not reflect natural emotion [335, 336].

[3]This concept is also described by Cowie [154] as effect- and cause-type studies

4.3.1 Simulated Vocal Expressions

In previous studies, the most common and easiest method of obtaining speech material is by asking professional or lay actors to produce vocal expressions of emotions [260, 189, 326, 337]. One of the notable benefits to this method is the additional experimental and environmental control. Since the acted emotion is predetermined, no perception tests are necessary to label the data. However, researchers can perform subsequent perception tests to verify that listeners actually perceive the portrayed emotion as intended. While portraying a given set of discrete emotions, the verbal content in a speech utterance can be controlled and standardised [12]. This way the perceived emotion is minimally influenced by the semantic channel, which allows the researcher to identify how well listeners decode emotion from the tone of the voice independently. Presuming that each acted emotional state is expressed exclusively, the researcher can make direct comparisons of acoustic content in phonetic differences, and attribute them to the associated paralinguistic information.

Simulated vocal expressions are mostly more intense than spontaneous emotional states [12], and likely to represent culturally shaped prototypical expressions [269]. This can be an advantage, yet also be a limitation. It is an advantage because it makes the expression easier to identify. Generally, acted material is easier to classify automatically compared to spontaneous speech [241, 193, 338] because the emotions portrayed produce higher arousal levels. However, intense prototypical emotions are not commonly found in day-to-day scenarios. Actors may fail to produce the more subtle cues, which may be the relevant ones, while at the same time over-emphasise others [73]. A major advantage is the fact that the environment is easy to control. This allows for good sound quality on the recordings, which is crucial for effective acoustic analysis, particularly at the early stages of research where many aspects of emotional communication are undetermined.

Although the use of simulated speech has numerous advantages, there are distinct limitations:

- simulated speech is often non-interactive [189, 115], hence may only provide for a lim-

ited range of emotions [339];

- emotion portrayals reflect conventionalised stereotypes [12] and may not include involuntary physiological responses to stimuli normally associated with emotion [340];
- actors may have different subjective interpretations of the emotions they are instructed to simulate, and acting (or encoding) abilities can differ considerably between actors [144];
- acted speech is often read from text, which differs significantly in acoustic characteristics due to its nature [341].

Some of these issues can be addressed. The quality and perceived naturalness of portrayals, for example, can be ascertained by conducting judgement studies [189]. Furthermore, stereotypical portrayals can be avoided, and interactive dialogues can be used. Elicitation techniques can be improved by modelling theatrical performance strategies [342].

4.3.2 Natural Vocal Expression

The ideal research paradigm seems to be that of natural vocal expressions. Hence, more recent studies have shown a tendency towards developing natural, real-life speech corpora [159, 343, 157]. Speech data containing natural vocal expressions have been sourced from various situations, such as dangerous aviation circumstances, therapy sessions, call centres, reality television, and journalist reports (see [241, 12, 201] for a list of studies). The primary advantage of natural spontaneous speech is the increased likelihood of having high ecological validity [201]. As mentioned above, in an ideal scenario the subject is in a natural setting, is unaware that the experiment is about emotion, and if practicable, can be recorded in an inconspicuous manner. Picard et al. [334] make an interesting point about the fact that recording devices are becoming increasingly more common, which may lead to the subjects becoming less aware of them.

Due to copyright restrictions, much of the broadcast material is not available for research. Privacy and ethical restrictions prevent covert recording of subjects to take place. For this reason,

truly natural material is generally difficult to obtain. Moreover, finding appropriate methods for labelling naturalistic data is more complex (section 3.3.4). Issues that came to light studying naturalistic data were not evident with that of acted material. Most full-blown, prototypical emotions are often absent in realistic databases [179]. Emotion representation for naturalistic speech, therefore, is compelled to depart from traditional uses of predefined lists such as the big 'n' emotions, which are seldom found in day-to-day speech. As a result, research is veering towards methods that allow the more subtle underlying emotions to be included, such as boredom, interest, etc. Subtle emotions are more difficult to recognise by humans, and therefore, the task for developing systems to successfully recognise spontaneous emotion will also be more difficult. To complicate matters still further, one cannot assume that natural vocal expressions represent emotions that are actually felt. Spontaneous speech communication is diffused with complex variability of involuntary and voluntary control, the distinction between expression of genuinely felt emotion and strategic signalling [29]. Some of the issues include regulation mechanisms such as masking, acting, and strategic deception, and regulations that may be culturally or socially defined, such as display and feeling rules. These conceptual influences have been referred to as 'push' effects [12, 340], and 'input-related issues' [19, 156] (see section 4.3.4).

Because natural data is mostly obtained in a natural, uncontrolled environment, it is difficult to determine the precise nature of the appraisal criteria that may have induced the emotional state, which is unfavourable for studies that incorporate cognitive models of emotion [29]. In addition, recorded material in an uncontrolled environment is more likely to be of poor sound quality—audio quality is an aspect that is often undervalued. Generally, real-world environments are noisy, and obtaining a clean signal of the voice in such situations is difficult. One must be aware that certain voice measures, which could be of significant relevance, are sensitive to extraneous environmental noise. In fact, some of these acoustic cues may be the exact ones that are particularly difficult for actors to produce [29].

4.3.3 Induced Vocal Expression

To find a balance between the limitations that come with acted and truly naturalistic data, a good trade-off between controllability and naturalness is to induce emotions in participants in a laboratory environment. Although such material is quite sparse, this method is becoming increasingly more attractive as a promising compromise [15, 194, 29, 12]. Effectively, investigators can use any one of a variety of induction techniques to trigger an emotional response in a subject, and subsequently record the associated vocal responses. Gerrards-Hesse et al. [344] provide a review of the effectiveness of Mood Induction Procedures (MIPs) used in nearly 250 studies. Although there are countless approaches for inducing emotion, they proposed that most MIPs fall under the following five groups:

1. *MIPs based on the free mental generation of emotional states:* In this group, the stimuli are activated by the subjects themselves, which involve emotion-inducing techniques such as mental imagery (Imagination MIP) [345] or hypnosis (Hypnosis MIP) [346].

2. *MIPs based on the guided mental generation of emotional states:* This group uses MIPs that present the subject with emotion inducing material with the additional instruction to get involved with the suggested mood state. These include the Velten MIP (subjects are asked to read statements that describe positive or negative self-evaluations) [347], the Film/Story MIP (subjects are asked to imagine the situation in the presented film or story) [348] and the Music MIP (subjects listen to a mood-suggestive piece of music) [349].

3. *MIPs based on the presentation of emotion-inducing material:* This group assumes that the presented emotional stimuli will induce an emotional response without the explicit guidance of getting into the suggested mood. It uses external material similar to that mentioned previously, with the additional Gift MIP—with the assumption that the participant with be elated with an unexpected gift [350].

4. *MIPs based on the presentation of need-related emotional situations:* MIPs in this group exploit the subject's susceptibility towards satisfaction and frustration, such as the need for achievement or affiliation. These include the Success/Failure MIP (giving the sub-

ject false-positive or false-negative feedback of performance during cognitive tasks) and the Social Interaction MIP (subjects are exposed to certain social interactions specifically arranged by the experimenter), which can be used to manipulate certain appraisal dimensions [351, 352].

5. *MIPs aiming at the generation of emotionally relevant physiological states:* These include MIPs that induce physiological arousal by administering psychoactive drugs or placebos (Drug MIP) [353]. In accordance with the facial feedback hypothesis, subjects are asked to contract or relax different facial muscles to produce a frown or smile (Facial MIP) [51].

Induction experiments are favoured because of the degree of control they provide. Many inducing techniques have, of course, ethical constraints. For example, administering psychoactive drugs in participants, is nowadays unlikely to get ethical clearance. Similarly, there are ethical constraints that prevent the induction of strong emotions in a laboratory environment, limited to providing relatively low-intensity emotions. This can make it difficult to differentiate between states [269]. There are, however, many studies that have successfully induced emotion to study the effects on the voice (e.g. [31, 159, 352, 354, 355, 356, 357]). To prevent demand effects[4], the experiments are in most cases designed so that subjects are unaware that the objective of the experiment concerns emotion. The Success/Failure MIP seems particularly apt in this regard [358] because the true nature of the experiment can be disguised (cf. [159, 352]). Because of the degree of control that MIPs provide makes them favourable for exploring cognitive models of emotions, such as the component process model by Scherer [151]—such a degree of control also favours work within the Brunswik lens model [29] (see section 4.1.1). In this framework, a researcher manipulates cognitive appraisal dimensions and subsequently measures any voice alterations, or physiological changes. It should be noted that inducing techniques can produce varying responses between different individuals [12]. For this reason, it is important to verify the emotional content by performing subsequent judgement analysis and verify consistency and reliability amongst listeners.

[4]Demand effects occur when the subject can guess the purpose of the procedure and hence act the desired emotion.

4.3.4 Authenticity related Considerations

Most would agree that the emotion expressed by a speaker, and subsequently observed by a listener, does not necessarily reflect the emotion that is truly felt. In fact, it is difficult to obtain a true evaluation of the speaker's subjective experience. It may, therefore, be assumed that an evaluator assesses emotional content of an utterance differently to how the emotion is experienced [152]. Although vocalisation is naturally responsive to emotional experience, one does have some voluntary control over it [359]. The distinction of involuntary physiological reactions and voluntary strategic signalling has been termed by Scherer as *push-* and *pull-effects*, respectively. Push-effects correspond to the involuntary physiological responses to emotions, such as muscle tension, increased heart rate and respiration, vocal tract changes, and facial expressions. These effects may cause emotions to 'leak' through, despite efforts to conceal them. Pull-effects[5], on the other hand, correspond to the voluntary control of emotional behaviour to comply with external conditions such as social and cultural norms, referred to by Ekman [360] as *display rules*, and/or the voluntary control of strategically misrepresenting or concealing one's emotions, e.g. *deception* [361, 362]—similar distinctions that differentiate genuinely felt emotion with strategic signalling have been made by other authors (see [29, p. 82]). On the one hand, Banse and Scherer [189] suggest it is unlikely that strategic or emotive communication would differ strongly in its signal patterns—and, therefore, argued for the validity of portrayed expressions—while on the other hand, Campbell [363] stressed that making such a distinction is significant. For acted speech, for example, although listeners may detect the intended emotion portrayals correctly, they may also be aware that the speaker is consciously intending to express such an emotion.

4.3.5 Audio Quality related Considerations

During earlier times, providing hardware and software for extensive acoustical analysis and manipulation was less accessible and more costly compared to nowadays. Instead of using expensive hardware, voice analysis is now done on basic computer software systems. Such

[5]This is similar to the concept of input-related issues [19, 156]

systems increase convenience. The criterion needed to obtain optimal recordings of speech is an aspect still often underestimated. It seems few examples in the literature provide details of the audio equipment used [364]. The desired signal can be affected by factors such as: the type of recording systems used; microphone type, placement, and angle; and unwanted environmental noises [331]. To prevent extraneous acoustic factors, a good recording environment should make use of sound-proof booths [365], which ultimately will minimise noise-induced errors during acoustical analysis. A common error caused by noise, for example, can result in F_0 estimates being off by an octave [15]. Furthermore, noise affects perturbation measures such as *jitter* and *shimmer*, which are also substantially influenced by microphone sensitivity and the distance placed from the sound source [366]. Even different microphone types have been shown to have a significant impact on speech parameter values [367]. Low-pass filtering is often used to remove a selected range of noise before analysis, but this only attenuates the frequencies above a given cutoff frequency. Noise reduction technology has improved remarkably in recent years. However, in most cases there are shared frequencies between noise and the acoustics of speech, hence the removal of noise will impair the speech signal to some extent. Not only do sound-proof booths isolate the desired sound source, i.e. the speaker, they also minimise reverberation. Reverberation can affect F_0 contour, temporal envelope, and formant transition measures, and it can increase low-energy levels and reduce amplitude modulations. In fact, it generally modifies the overall timbre of the speech signal [368].

4.4 Discussion

This chapter of the review identified several major research topics in the area of vocal communication of emotion. First, we introduced Scherer's revised Brunswikian lens model, a framework that describes the complete process of emotional communication through speech (section 4.1.1). It has been suggested that the model is methodologically highly suitable for the study of vocal expression [29, 12]. One important aspect it highlights is that the perceptual representation (proximal indicators) of emotion does not necessarily correspond in a one-to-one fashion to the objectively measured cues (distal indicators) captured at the encoding stage [12]. The

framework distinguishes between studies of encoding, decoding, transmission, representation, and inference. It considers several different aspects that may influence the uncertain relationship between encoders, cues, and decoders. According to the issues identified by the model, the potential study types in the area of vocal communication of emotion were differentiated and discussed (sections 4.1.2–4.1.6), providing us with both research objectives and direction. Although in this thesis the Brunswikian lens model is not implemented in its entirety, there are many aspects of our work that fit nicely into this framework.

As outlined in section 4.1.2, a typical encoding study is typified by the nature of the material investigated, which is commonly distinguished by emotion elicitation type: simulated (acted), naturalistic, and induced emotion (section 4.3). Common discussions about the distinctions of these data types include emotion authenticity/genuineness, ecological validity, and audio quality. Up to about a decade ago, many were debating whether the essence of acted material could be representative of natural emotions. Nowadays, it is widely acknowledged that acted data cannot be adequately compared with naturalistic data [369, 332, 370, 338, 313]. However, despite acted speech being radically different in nature, its relevance to perception is not completely unrelated. Much knowledge to date can be attributed to studies of simulated material. In fact, acted speech material is still widely used (e.g. [371, 372]), but mostly the findings are not associated with data encountered in real-life situations. The distinction lies with the application in mind. For example, acted data can be appropriate for applications potentially found within the entertainment industries, such as animated characters in computer gaming and movies, while naturalistic data can be appropriate for real-life applications such as medical diagnostics.

Increasingly, studies are showing a tendency towards developing natural, real-life speech corpora [159, 343, 157] (with less emphasis on full-blown emotions). Compared to simulated emotional speech, work on naturalistic speech has proven more complex for several reasons. These include factors such as the acquisition of suitable data (section 4.3), choice of appropriate emotion descriptors (Chapter 3), the establishment of actual felt or observable emotional states

(section 3.2.5), and effects such as the push and pull distinction, display rules, and strategic signalling (section 4.3.4). Obtaining well founded naturalistic emotional speech data is not a straightforward task. The previous work undertaken, which provided the speech material for this thesis [159, 339, 373], focused on natural speech and addressed two pressing matters that can be seen as relevant to the Brunswikian framework. First, one of the objectives of their work was to provide emotional speech recordings of the highest audio quality. Prioritising speech signal preservation (transmission) minimises any intervening variables that may systematically alter the distal cues in the transmission process (section 4.1.5). For example, the use of sound-proof booths can minimise any extraneous acoustic factors such as reverb, which can have an impact on fundamental frequency (F_0) estimates. Second, the authenticity of the emotion present (encoding) in the speech material is central to defining its type (naturalistic), as outlined in section 4.3. As the issues with both acted and truly naturalistic data became evident, Cullen et al. [364] turned their focus towards inducing emotions (section 4.3.3), arguing that these methods are an appropriate compromise between the factors associated with them. Their work implemented mood inducing procedures in a controlled laboratory environment, and delivered high-quality speech material that reflects natural expressions. The accompanying challenge with building a corpus involves developing suitable methods for labelling the speech material that describes its emotional content (decoding studies). In the last chapter we mentioned that the suitability of labelling methods are partly contingent on the eliciting type present in the material, i.e. whether the emotions are simulated or spontaneous. As Ellen Douglas-Cowie explains, both strands, collecting speech data and labelling it, are very much interconnected, and says that "the categories used in labelling are driven by the material that is there to be labelled, and the collection of material is driven by an understanding of the categories that it is relevant to collect." [374]. As alluded to in the previous chapter, it was decided to carry out a case study to label the given naturalistic speech material, which will finalise the construction of the naturalistic emotional speech corpus (Chapter 7). This case study will serve to explore the previously proposed research question regarding the appropriateness of the labelling scheme (RQ2). In addition, the ratings obtained from this case study will allow us to verify the emotional content present in this given speech dataset, essentially allowing us to examine the effectiveness

of Mood Inducing Procedures—although there is a mutual dependency between verifying the appropriateness of the descriptive scheme and verifying the emotional content. To this end, this chapter has given rise to the following research question:

> **RQ3:** Can mood induction procedures provide naturalistic speech with sufficiently discernible levels of emotion?

There is ample evidence to show that listeners can reliably infer affect-related arousal from vocal acoustics, and that specific acoustic features associate with the perceived arousal [42, 12, 141]. However, although listeners can reliably differentiate between emotions, hedonic valence remains, to date, to be more difficult to detect from unique acoustical patterns, both in automatic recognition (building machine learning classifiers) and by human listeners [253, 375, 235, 376, 12, 139, 13, 377]. There are several reasons that may explain this:

1. The incorrect or incomplete selection of acoustic cues is being studied [29] (discussed further in the next chapter).
2. Probable arousal differences between emotion families are not being considered adequately [12] (section 3.3.1).
3. Acoustic correlates of emotion are studied when judgements are not made solely on acoustic information [378, p. 30]—neglecting the interaction with verbal meaning (section 4.2).

We suggest that the latter reason is a very likely candidate. This notion was discussed in section 4.2, where we pointed out that labels are often provided based on judgements made on both semantic and acoustic content of speech. For natural spontaneous speech, it is difficult to isolate either the semantic or acoustic channel so as to investigate it independently. Several techniques are available that attempt to address this (section 4.2.2). By manipulating certain cues to mask the semantic content (section 4.1.4), the method that we emphasised was low-pass filtering. This method has been used in previous studies and seems to be a useful tool to remove semantic content. To examine the extent to which labels are based on acoustic cues exclusively, we suggest to low-pass filter speech to explore the following research questions:

RQ4: Does nonverbal naturalistic speech convey Activity and Evaluation levels that are recognisable to listeners?

RQ5: How do ratings from two perceptually different conditions (verbal and nonverbal speech) compare?

4.5 Conclusion

This chapter has given us additional knowledge on the labelling of emotional speech, provided the conceptual basis of the vocal communication of emotion (sections 4.1 and 4.2), and introduced the available speech data sources that a study can be typified by (section 4.3). Outlining the different processes involved with the vocal communication of emotion (section 4.1) gives us a better understanding of what is involved in labelling emotional speech, contributing to research question two from the previous chapter (RQ2). Subsequently, research question three (RQ3) seeks to examine the potential use of Mood Inducing Procedures for delivering naturalistic emotional speech corpora (section 4.3.3) by examining the emotional content present. To this effect, we investigate whether listeners are able to decode expressed levels of Activity and Evaluation (two-dimensional model of emotion) in the given MIP-based speech material. This work is pursued in the case study reported on in the next chapter (Chapter 7). With regard to the information presented in sections 4.1 and 4.2, the concept of different cues (and channels) affecting emotional decoding, prompted the idea of investigating the extent to which acoustic cues determine the perception of emotion dimensions (activation and evaluation) in mood induced speech (RQ4 & RQ5). Effectively, this allows us to determine the validity of current labelling methods that are based on both acoustic and semantic information. To this end, we will investigate how listeners utilise voice cues to infer emotion from low-pass filtered speech (cue masking, section 4.1.4), which isolates prosodic features in lower frequency regions, while making speech unintelligible. This work will be presented in Chapter 8.

5

Acoustic Correlates of Emotion in Speech

Vocalisations are produced from physiological changes that include respiration, phonation, and articulation. There is a considerable amount of evidence to suggest that emotions modulate these physiological changes, which in turn partly determine the produced acoustic signal [189, 379]. These notable variations occur at the suprasegmental level of speech communication and can carry a large amount of nonverbal information, including the emotional state of the speaker [141]. Both acoustic and linguistic features can be extracted from the raw speech wave data to provide information about emotional states [10], although acoustic features are the more typical features used so far [296, 297, 298, 280, 299, 300, 301, 302, 303, 304, 305, 306]. Research in this area is still ongoing as there is currently no general agreement on which features are the most important. Voice cues can be broadly grouped into categories related to fundamental frequency (e.g. pitch contours), time-related measures (e.g. speech rate), intensity-

related measures (e.g. signal energy), voice quality (e.g. jitter and shimmer), and combined time-frequency-energy measures (e.g. long-term average spectrum). The first three categories mainly represent perceptual dimensions of pitch, speech rate, loudness, phonation type, respectively, whereas combined time-frequency-energy measures illustrate perceived timbre [269]. According to Scherer [12], most studies have limited their measurements to F_0, energy, and speech rate parameters. He suggests that these measurements are more likely to reflect the speaker's arousal (or activation) state, rather than hedonic valence differences. Instead, source and articulation characteristics such as frequency distribution and formant parameters may be more indicative of qualitative valence differences. In this section, we present a non-exhaustive overview of some of the more classical description of speech acoustics often used in the use of emotion in speech studies.

5.1 Source-filter Theory

The description of vocal acoustics as a two-stage process involving a combination of source energy and resonance effects is central to the *source-filter* model [380]. This concept was originally developed to account for the acoustics of vowels but is now routinely used in other areas of speech analysis and synthesis [15]. The principles of the source-filter model are also considered when analysing speech acoustics for emotion [13, 15, 73]. Both source- and filter-related cues are relevant indicators of physiological changes in vocal production that may accompany emotional arousal [379].

As the air passes from the lungs through the *glottis* it causes the *vocal folds* (located in the cartilaginous *larynx*) to vibrate, creating source energy. The source energy is subsequently modified (filtered) as it passes through the *supralaryngeal vocal tract*, which comprises the pharyngeal, oral, and nasal cavities (see Figure 5.1). The source energy can be either *voiced* or *unvoiced*. The regular vibration of the vocal folds produces *voiced* (or phonated) sounds such as vowels [15]. The resulting speech waveform is then *quasi-periodic*[1]. The basic rate at which the vocal

[1]Quasi-periodic means that the speech waveform shows periodicity over a short-time period (5-100 ms) during

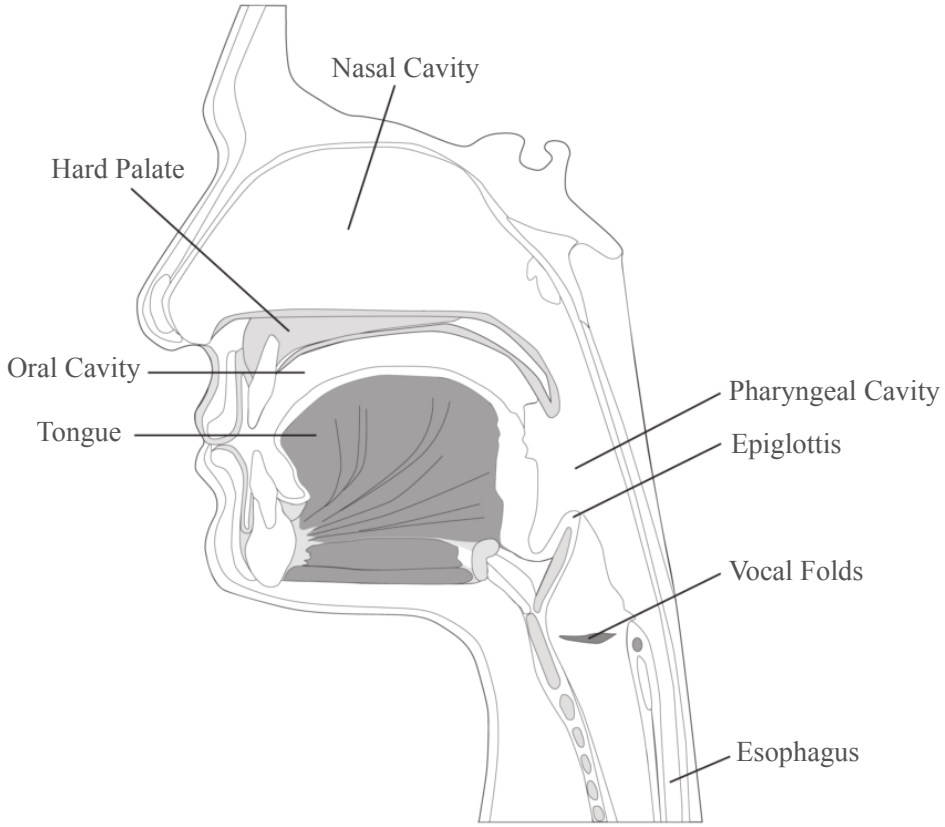


Figure 5.1: A schematic illustration of the human vocal tract [13].

folds vibrate is called the "fundamental frequency" (F_0), which corresponds to what a listener perceives as *pitch*, i.e. the part of speech that gives it the tonal quality. The size of the vocal folds (length and mass) and the tension placed on them determines the rate of vibration [382]. Most adult males, for example, have longer vocal folds that produce lower vibration rates, perceived as lower pitched voices. Most females, and particularly children, have shorter ones that produce higher pitched voices. Unvoiced sounds, in contrast, are produced when the air passes through the glottis without causing the vocal folds to vibrate. The resulting speech signal is subsequently voiceless, generally perceived as noisy and breathy (e.g. "h" and "s" sounds). The waveform that is produced is then *aperiodic* (i.e. random noise).

The energy leaving the larynx passes through the supralaryngeal vocal tract, where the resonance properties of the pharyngeal, oral, and nasal cavities act as filters. These acoustic differences result from energy regions being passed or attenuated, creating high-amplitude energy

which it is stationary [381]

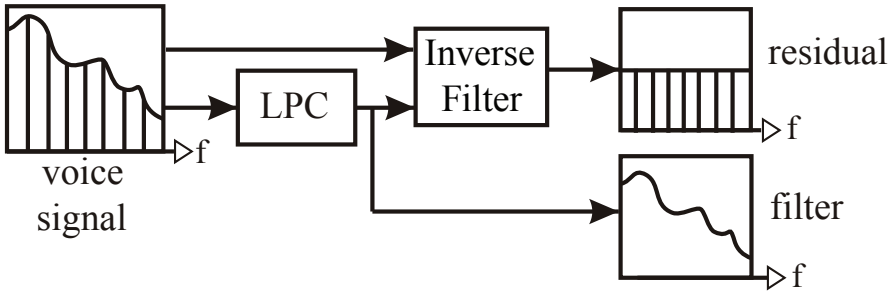


Figure 5.2: Diagram illustrating a linear prediction coding (LPC) analysis algorithm. The LPC residual has a flat spectrum as a result of minimising the error between the signal's spectrum and the frequency response of the filter [14].

bands at resonance locations known as *formants*. The frequencies that are amplified or attenuated are determined by the characteristics of the vocal tract such as the vocal-tract length, position of the tongue, lips, jaws, etc. The ability to make rapid movements with the tongue and jaw to change the size and shape of different cavities allows us to produce articulated speech. Even facial expressions, such as smiling, can significantly affect the outcome of vocal tract resonances, specifically reflected in *formant* frequencies [13]. Such filter-related cues may, therefore, be important indicators of expressed emotion but seemed to have not been studied by researchers in the field so far [42, 12].

To provide a perceptual approximation of the glottal sound source one can apply an inverse filter (IF) to the original signal. The most common method for doing this is linear prediction coding (LPC). An LPC algorithm (see Figure 5.2) finds a filter to fit the spectrum of the input signal, applies the inverse, and extracts the LPC residual (glottal source)—numerous automatic inverse filtering algorithms have been suggested (see [383] for an overview). LPC models the glottal source with a fixed spectral envelope, although the spectral envelope of the actual voice source generally varies. In fact, the spectral envelope is shaped according to different voice qualities, such as vocal effort and lax voice [14]. Voice quality is a source-related cue, which corresponds to the variability in the frequency and amplitude of vocal-fold vibrations (see section 5.6).

5.2 Prosody

In a general sense, prosody describes the way one vocalises a sentence. It refers to the temporal and melodic aspects of a spoken language that give a sentence meaning beyond its lexical content. Prosody is at the *suprasegmental* level of communication, i.e. the phenomena that spans speech segments (phones) [384, 385]. The suprasegmental features that collectively combine to form prosody in speech production and perception include *rhythm*, *stress*, and *intonation*. A simple example of prosodic variation is the difference between a declarative statement ("They are gone.") and a question ("They are gone?"). In this example, pitch variation (usually with a rise at the end of the sentence) is used to convey non-lexical information about whether the sentence was a question or not. In tone languages, such as Chinese-Mandarin, melodic oppositions have phonemic value that can distinguish words from one another [384]. In most European languages, however, prosodic features do not typically change the meaning of a word [386].

Cutler et al. [387] point out that the term prosody is used in different ways by different researchers. Some conceptualise prosody as an abstract meaning of the structure of speech, while others conceptualise it more at the actual acoustical conveyance, effectively as a synonym for suprasegmental features, such as pitch, tempo, and loudness. In fact, Werner and Keller [384] distinguish between four different representations of prosodic conceptualisation (cf. [385]):

- **The linguistic intention** is the use of prosody to make semantic distinctions such as the question-statement distinction mentioned above. Prosody also accentuates certain elements of a text by marking boundaries and conveying transitions between words, phrases or sentences. The linguistic representation refers to prosodic descriptions relating to *tone*, *intonation* and *stress*.

- **The articulatory manifestation** is the systematic modifications of articulatory movements that produce prosodic variation. These observable physical movements include variations in *amplitude of articulatory movements*, *air pressure*, or *electric impulses* in the articulatory musculature.

- **The acoustic realisation** of prosody refers to the variables that can be measured from an acoustic signal. The main acoustic variations include *fundamental frequency (F_0)*, *intensity* and *duration*. A stressed syllable, for example, is often higher in fundamental frequency, greater in amplitude, and longer in duration compared to unstressed syllables.
- **The perceptual representation** refers to the human perceptual processing. At the subjective level of perception, prosodic phenomena include *pauses, length, pitch/melody* and *loudness*.

It is generally acknowledged that the main prosodic measurements of an utterance are variances of timing, amplitude, and frequency. Although, these features are, in fact, the measurable dimensions of sound itself. Prosodic features can be analysed over a syllable, word, phrase or an entire utterance (or turn), which makes it context sensitive. Moreover, various aspects of prosody can have duration relevance, e.g. vowel, pitch, and intensity durations. Similarly, stress can be distinguished over a word, phrase, or sentence, but is mostly conveyed on a single syllable [384].

Prosodic features are the most commonly used features for emotion in speech recognition [152, 388, 19, 303]. Murray and Arnott [389] suggest that "anger" and "joy" have a faster speech rate, a higher pitch average, a wider pitch range, and higher intensity. Sadness and boredom, on the other hand, are characterised by a lower pitch average, slightly narrower pitch range, and a slower speaking rate. Since anger and joy have been shown to share the same vocal characteristics, it seems that prosody characterises activation rather than its valence, since anger and joy are opposites in hedonic valence. Fragopanagos and Taylor [19] point out that the findings from most prosodic studies of emotion so far tend to be limited to information about the activation level, rather than the qualitative valence of an emotional state.

Prosody Frameworks and Applications

One of the most popular labelling schemes to represent prosodic events for categorical annotation is the Tones and Break Indices (ToBI) framework [390]. ToBI is a framework for

transcribing the intonation and prosodic structure of spoken utterances. This framework measures prosodic features such as pitch accents (or prominence) and prosodic phrase boundaries (the perceived grouping of words in an utterance) [391]. It comprises a series of labelling tiers, a tonal tier, break index tier, and a miscellaneous tier. Similarly, LinguaTag uses a three-tiered approach to define stress in prosodic events, but it consists of duration, pitch and intensity [392]. Other intonation models include IPO [393] and Tilt [394]. The IPO (Institute of Perception Research) approach was originally for the research of Dutch intonation but has also been adapted and applied to other languages. The approach provides a framework that extracts raw acoustic F_0 data to provide a model of intonation for a given language. The intonation is represented by a series of discrete pitch movements (rather than pitch levels) for which the standardised curve, when resynthesised, should be perceptually equivalent to the original contour [395]. Mozziconacci and Hermes [283] investigated the production and perception of utterances expressed with emotion using the IPO model. By manipulating cues via synthesis (as mentioned in 4.1.4), they examined the perception of emotion in speech by transferring a series of different intonation patterns (obtained from acted speech) onto neutral utterances. Similarly, the Tilt intonation model facilitates automatic analysis and synthesis of intonation. It provides a parameterised representation of the intonational events in F_0 contours. Pitch events occur as instants in a linear fashion, have distinct start and end points, and detail pitch peaks and throughs [395]. McGilloway et al. [303] used a system called ASSESS (Automatic Statistical Summary of Elementary Speech Structures) for preprocessing to retrieve measures from the speech signal. In their study, prosody was the primary supra-segmental feature they used for statistical analysis.

5.3 Fundamental Frequency (F_0) related Measures

As mentioned, voiced sounds are the result of the vibration of the speaker's vocal cords. The rate at which the vibration occurs is defined by the quasi-periodic number of cycles per second, measured in hertz (Hz). This is called the *fundamental frequency (F_0)*, and its relative perceptual impression is referred to as *pitch*. According to the American National Standards

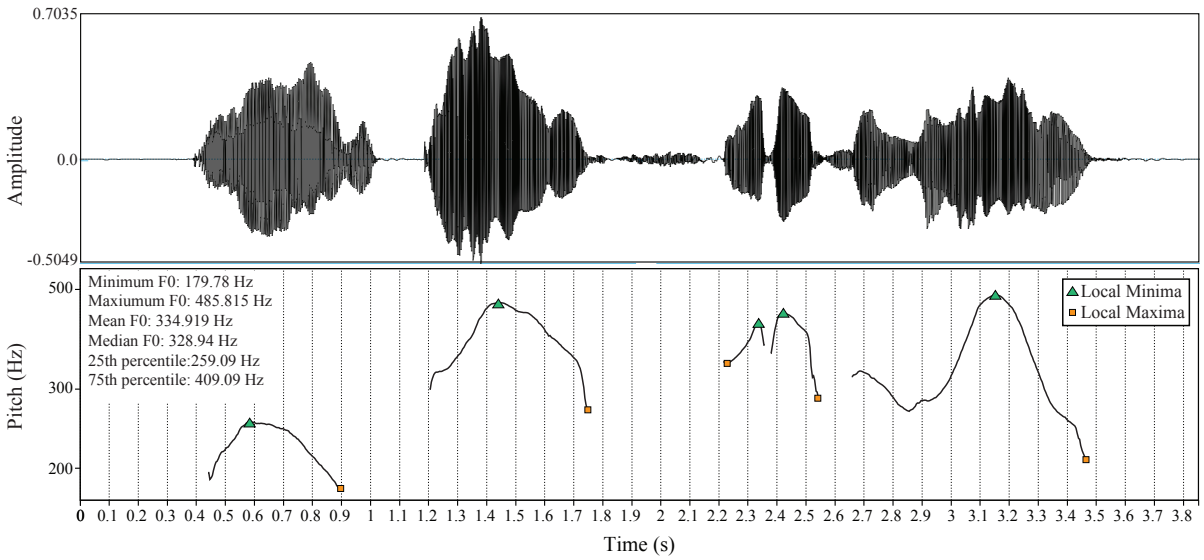


Figure 5.3: Speech signal (top) of the utterance *we were doing so well*, spoken by a female speaker, in the time domain with its corresponding pitch contour (below).

Institute (ANSI) of 1994, pitch is defined as that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Pitch is that sensory (subjective) attribute to melody, harmony, and tonality. Intonation plays a big part in natural speech and is primarily related to fundamental frequency (F_0) [384]. Intonation involves incidences of recurring F_0 patterns that are represented by the *pitch contour* [396]. The pitch contour characterises the temporal evolution of pitch within a speech utterance [296]. This feature is an important aspect of emotion in speech recognition, as it can describe the temporal characteristics of an emotion as it unfolds in time [241]. The properties of a speech wave form are commonly illustrated as periodic variation of sound pressure amplitude as a function of time. Waveforms are typically illustrated in a *time domain*. Figure 5.3 shows a speech clip illustrated in the time domain (top) with its corresponding pitch contour (bottom) over time.

Common pitch statistical measurements of speech include mean, median, minimum, maximum, standard deviation, range, first quartiles, third quartiles, and interquartile range. Similarly, one may come across the terms *topline* and *baseline*, which are effectively the trend of pitch peaks and throughs over a phrase [397]. Pitch extraction is a common procedure in the work of speech science, and is found in most speech analysis software. However, results may vary as there are

different algorithms for pitch extraction [15]. Typically, algorithms for pitch extraction correlate the waveform against itself to find repeating patterns of periodicity in the signal, and then producing an autocorrelation function on it. Errors can occur due to noise, present due to unwanted signals in the recording, or merely as part of the sound produced in the speech production system. The most common errors found are "octave jumps", particularly for emotional speech [30, p. 235]. These errors are often manually corrected.

Pitch is an important component for conveying non-textual information, such as prominence and emotion expression [141]. This seems to be the case for many languages [115]. Ververidis and Kotropoulos [241], for example, showed that anger had the highest pitch level, followed by fear. Disgust, on the other hand, is expressed with a low mean pitch level. They noted that the majority of research reported a wide pitch range for fear. Juslin and Laukka [242] provide an extensive review of the findings of F_0 correlates of emotion in 104 studies. In most reviewed studies, F_0 level was found to be high for anger, fear, and happiness, and low for sadness and tenderness. F_0 variability was found to be high for anger and happiness, but low for fear, sadness, and tenderness. Meanwhile anger, fear, and happiness were associated with a higher proportion of upward F_0 contours compared with sadness and tenderness. With respect to frequency, intensity, and duration, they found that positive emotions exhibited more regular patterns compared to the irregularities found in negative emotions. Furthermore, they noted that much reviewed work was based on data from "informal observations or simple acoustic indices that do not capture the complex nature of F_0 contours in vocal expression". They suggested that further research is needed to confirm these preliminary results.

5.4 Time-related Measures

As an integral part of the prosodic phenomenon, the variations in the speed of speech production play a big role in the expression of emotion. These variations correspond to changes in *speech rhythm* or *speech rate* [384]. There are several methods for studying speech rhythm [398], yet no standardised measure exists. Generally, speech rhythm corresponds to the dura-

tions of vocalic and consonantal stretches in a speech signal. It can, for example, be measured over vowel duration [399]. Speech rate, on the other hand, is the amount of speech produced over time, normally measured by quantifying the number of phonemes, syllables, or words in an utterance per second [15]—syllable counts being the most common method. Speech rate measurements generally include speech pauses, whereas calculations based on speaking time excluding pauses have been referred to as "articulation rate" [400]. In itself, variability of pauses can also be potential indicators of emotional expression. Other temporal measures have included mean of silence duration, mean of syllable duration [28], and the inverse of mean length of voiced parts [401].

The review provided by Juslin and Laukka [242], suggests that speech rate/tempo is of primary importance for listener judgement on vocal emotional expression. For example, they established that anger, fear, and happiness had a fast speech rate/tempo, while sadness was associated with a slow speech rate. However, not only does speech rate seem to be associated with more than one emotion, it is similarly associated with emotions in contrast of qualitative valence, such as anger and happiness. Moreover, it seems that there are gender-specific differences in speech rates associated with a particular emotion. For example, it has been reported that for anger males exert a slow speech rate, whereas females exert a fast speech rate [241, 402, 403]. A study by Laukka et al. [28] compared the general findings obtained from posed expressions with those obtained in their study of spontaneous speech. They found that the acoustic correlates of authentic irritation and resignation demonstrated a similar trend to those of posed anger and sadness (except for speech rate for irritation). They suggested that posed expressions, therefore, do reflect (at least partly) the physiological responses associated with spontaneous expressions. They reported, however, that the effect sizes for the correlates they obtained are much smaller compared to studies of posed expressions.

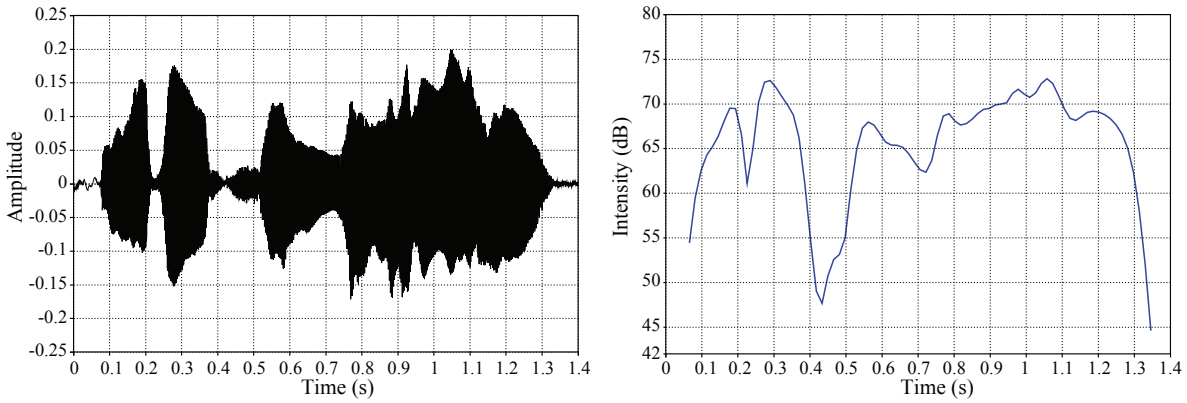


Figure 5.4: Speech signal of the utterance *we were doing so well* (left), with corresponding intensity contour (right).

5.5 Intensity-related Measures

Amplitude of a speech signal is a direct measure of the degree of displacement in atmospheric pressure caused by sound waves. Although it is directly related to the intensity of a sound, intensity is a measure that better reflects the perception and production of speech. Intensity is a measure of the amount of energy in the acoustic signal, measured in decibels (dB), which correlates with the perceived loudness and reflects the effort required in speech production [269]. A graphical representation of intensity contours can be illustrated (see Figure 5.4), similar to the illustration for fundamental frequency contours. Common global statistical measurements include mean, standard deviation, range, maximum, minimum, median, and mode. Intensity and variability of intensity have been shown [241, 242] to be important indicators of expressed emotion, particularly related to speaker arousal (activity). However, measuring intensity of speech is susceptible to several factors. These include the distance and angle placement of the microphone relative to the speaker, background noise levels of the recording environment, and recorder input levels [15]. To successfully calibrate measures of amplitude or intensity, cautious procedures when recording should be undertaken. In a natural setting, however, this is difficult to manage, so a controlled recording environment is more favourable in this regard. However, this will reflect on the emotion elicitation type found in the data, as either simulated or induced emotional expressions (see section 4.3).

According to the review by Juslin and Laukka [242], a high voice intensity mean is associated with anger, happiness, fear. Although, fear seems to be less reliable because a comparable number of studies showed that it was associated with a low voice intensity mean [241]. In contrast, sadness and tenderness were shown to have a low intensity mean. Similarly, voice intensity variability was high for anger, fear, and happiness, but low for sadness. In most cases, there seems to be a positive relationship between high pitch mean and high intensity. It may be that high pitch and intensity means correspond to emotions of a similar level of arousal [377]. In fact, the literature suggests that arousal (or activation) level is positively correlated with mean F_0, mean intensity, and, in most cases, speech rate [287].

5.6 Voice Quality

Like prosody, voice quality (phonation types) constitutes the paralinguistic form of communication. Voice quality is sometimes associated with laryngeal qualities but is mostly described as an auditory judgement of specific phonation types such as “tense”, “harsh”, “rough”, “bright”, “coarse”, “breathy”, etc. Voice quality can be conceptualised as part of a multi-layered prosodic system, distinct from the linguistic-prosodic pattern that carries semantic information [404]. Laver [405] defines voice quality as “the characteristic auditory colouring of an individual speaker’s voice”. In other words, voice quality can be described as the timbre of the voice. Voice quality is an important aspect in communicating paralinguistic information, such as individual identity, and expression of attitude and emotion [258, 406, 156].

Vocal quality parameters are determined by the different vocal fold vibration patterns, although they may not manifest immediate parameters as prosody does [258]. The vocal tract configurations are affected by various physiological states, which, in turn, do not sustain a constant period of oscillation. These variations affect voice timbres. This gives rise to two of the most common perturbations: *jitter* and *shimmer*. Jitter is the cycle-to-cycle variation of the period length (irregularities of pitch). Shimmer, on the other hand, is the cycle-to-cycle variation of the peak or average amplitude (irregularities in the intensity) [407, 258]. Jitter and shimmer

characterise voice qualities such as *roughness* and *breathy* speech [258]. Another feature associated with voice quality is the Harmonic-to-Noise (HNR) ratio. By measuring the degree of periodicity of a sound—the relative height of the maximum of the autocorrelation function [408]—HNR is a measure that quantifies (in terms of dB) the amount of additive noise (aperiodic) against harmonic (periodic) levels in the voiced signal. As mentioned previously, noise results from turbulent airflow generated at the glottis during phonation, due to inadequate closure of the vocal folds [409]. Noise may also be produced when the vocal fold vibration is aperiodic.

Several studies suggest that voice quality is fundamental in the expression of emotions [410, 383, 411, 121, 406, 412]. A study by Gobl and Ní Chasaide [383] focused on defining a certain relationship between voice quality and emotional states. The presented synthesised stimuli (see section 4.1.4) were obtained using resynthesis with a formant speech synthesiser system [282], which is an approximation of the intended voice qualities. Their results showed that there was no direct association between an individual voice quality and any one emotion (affect) [282, 383]. Moreover, Laver [405] associated "breathy voice" with intimacy, whereas "lax voice" (breathy voice at the phonatory level) has also been shown to correspond to sadness [73, 413]. Laukkanen et al. [413] reported that "breathy voice" was more indicative of sadness, surprise and enthusiasm. Gobl and Ní Chasaide [383] found that "breathy" stimuli did receive some "sad" and "intimacy" response but was more effectively signaled by the "lax-creaky" stimulus. For breathy voice they found no association with anger and happiness, although this contrasted with the findings of Murray and Arnott [389]. They mentioned that sadness was associated with a 'resonant' voice quality. Burkhardt and Sendlmeier [414] found a weak association with a breathy and 'falsetto' voice. A more general association of a tense voice might be with anger [73, 413, 414, 383]. Although, Scherer [73] suggests that tense voice is associated with anger, but also with joy and fear.

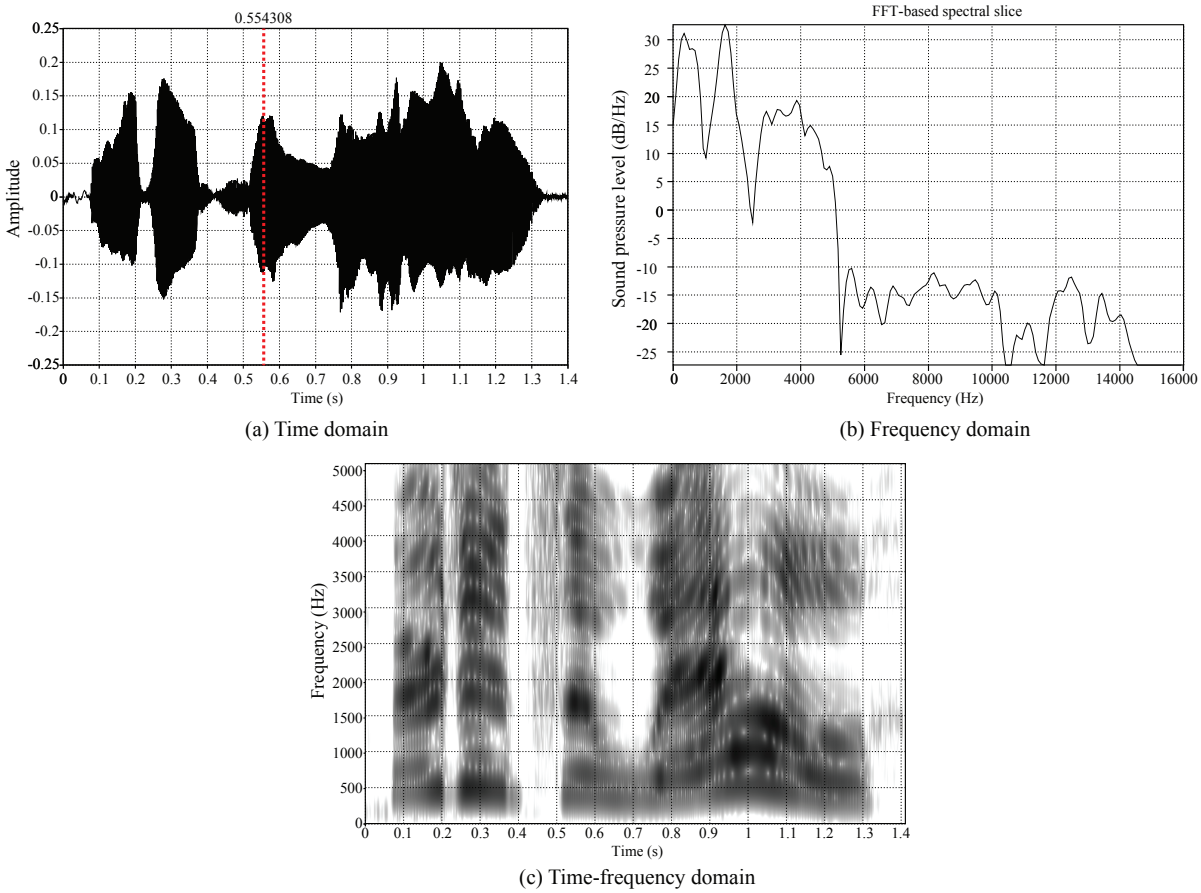
Gobl and Ní Chasaide [383] reported from their experiments that voice quality might provide a better indication of milder affective states rather than signalling strong emotions (except

anger). They note the difficulty in synthesising voice qualities, such as whispery, breathy and harsh stimuli, suggesting that one should be cautious when interpreting the results. Although voice quality seems to play a significant role in the communication of a speaker's emotional state, it is generally not used in isolation [383, 87]. It has been suggested elsewhere that voice quality is used to differentiate between discrete emotions [389, 73]. In terms of emotion dimensions, it has been argued that voice quality is a better indicator of emotional valence (e.g. [413]), while others argue that it is a better indicator of emotional arousal/activation and power/potency [383, 87].

The findings highlight the difficulty in comparing voice quality results. Most work on voice quality depends on the use of subjective auditory labels such as breathy, tense, etc. The problem with the use of such labels is that there is no real consensus between the descriptors used and the acoustic parameters measured, with results depending on the participant's interpretation of the terms used [73, 383].

5.7 Spectral Features

As mentioned in section 5.3, variations of pitch and amplitude are represented in the *time domain*. In speech analysis, it is common to transform the speech signal between the time domain (Figure 5.5 (a)) and the frequency domain (Figure 5.5 (b)) by means of a mathematical tool called the *Fourier transform*. Fourier analysis transforms a periodic signal to a function in the frequency domain, which represents the amplitude at each frequency. The transform is based on a theorem that proposes that any periodic time-series signal (time domain) can be decomposed as the sum of a set of sinusoidal functions (frequency domain), mathematically represented by sines and cosines. A Fourier analysis can illustrate both a single *spectral slice* and a *spectrogram*. The spectral slice displays the energy of each frequency component at a particular time, as depicted in Figure 5.5 (b) (at 0.554 seconds), or over a short segment, and thus has no time dimension. A spectrogram (Figure 5.5 (c)), on the other hand, visually represents the amount of energy at each frequency component as it varies with time. The



(a) Time domain

(b) Frequency domain

(c) Time-frequency domain

Figure 5.5: Speech signal illustrated in the time domain (a), as a single FFT-based spectral slice taken at 0.554 seconds represented in the frequency domain (b), and the time-frequency representation as a spectrogram.

amount of energy at a frequency component is represented by the intensity (or colour) at a particular point in time. In the frequency domain, the spectrum can depict features such as *harmonics*, *formants*, and *energy distribution*. To describe the spectral distribution, one can use global measures such as *spectral tilt*, *mean*, *standard deviation*, *skewness*, and *kurtosis* [15]. Another method to characterise energy distribution is to compute and subdivide the mean spectral energy over a smaller number of equally wide *frequency bands*.

Harmonics

Measuring the number of harmonics is a useful feature for speech emotion recognition [415, 241, 416]. Harmonics are component frequencies that are integer multiples of the fundamental frequency and are defined by their frequency and their amplitude [10]. The relative amplitudes

of the harmonics give characteristics such as voice quality beyond just pitch. Source variation and vocal tract filtering vary the amplitudes of the harmonics of a spectrum resulting in different voice qualities [417]. Vocal effort, for example, has more high-frequency content whereas a relaxed voice (or lax voice) has much stronger lower harmonics relative to the upper harmonics [14]. Comparisons can be made between the amplitudes of different harmonics, such as the first and second harmonics (H1-H2) or first harmonic and the strongest harmonic in the third formant (H1-A3). Epstein [417] states, for example, that a breathy voice usually has the highest amplitude in the first harmonic, and a creaky voice has higher amplitudes in the higher frequency harmonics. One should perhaps note that the perceived loudness of a voiced speech signal (proximal cues in terms of the Brunswikian framework) is not so much correlated with its overall intensity (distal cues) but rather with the amplitude of a few or even a single harmonic [12].

Formants

Formants are created by vocal tract resonances that amplify or attenuate certain frequencies in the spectrum. Formants allow us to quantify the natural shape and physical dimensions of the vocal tract. Formants are specified by their centre frequency, amplitude, and bandwidth. They are indexed above F_0 (fundamental frequency). The first formant is indexed as F_1, the second as F_2, and so forth [386] (see Figure 5.6). Voiced phones have four or more formants. In general, F_1 and F_2 are considered most significant to distinguish the phonetic properties of speech sounds, particularly for vowels [10], whereas the higher formants may be speaker dependent [405]. Furthermore, F_1 and F_2 appear to be more affected by emotional states than the other formants [241]. A common technique used to estimate formants is *linear predictive coding (LPC)* (see 5.1). This technique creates a spectral representation of time-series data by representing the most prominent spectral peaks using a small set of polynomial coefficients to define its function [15], effectively representing the filtering effects of the vocal tract.

Formants have been used in several studies in speech emotion recognition [418, 419, 420, 421,

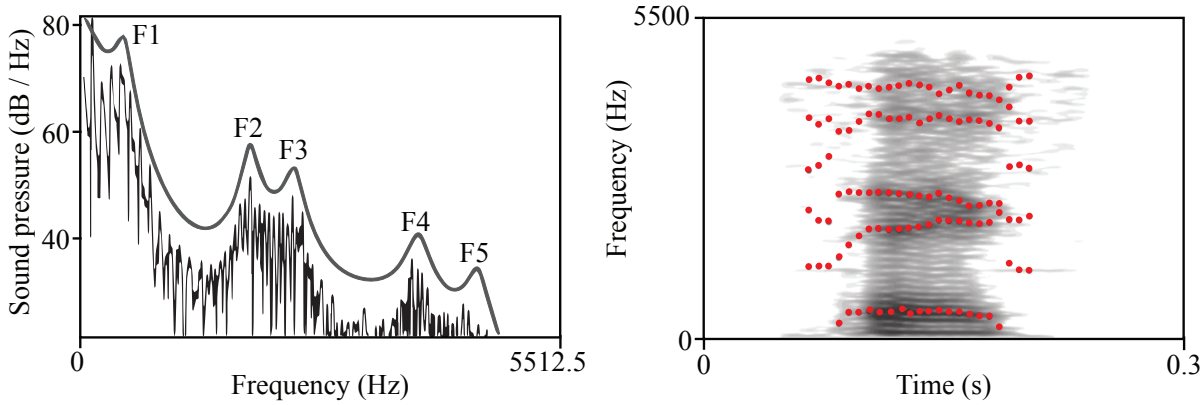


Figure 5.6: First five labelled formants overlying a Fourier spectrum. Linear predictive coding (LPC) was used for the envelope [15].

422, 423, 31]. Formants can be used to discriminate between different efforts in articulation that might be associated with certain emotions. For example, Ververidis and Kotropoulos [241] note that during slackened articulated speech formant bandwidth is gradual, whereas with improved articulated speech the formant bandwidth is narrow with steep flanks. Juslin and Laukka [242] found that formant measurements gave the most consistent results with anger and sadness, which were associated with increases and decreases in precision of articulation, respectively.

Spectral Tilt

The relative energy of higher harmonics with fundamental frequency (F_0) decreases over the frequency range. The slope of this harmonic spectrum is called the *spectral tilt* [276]. This is obtained by using linear regression to fit a line to the individual points that form the spectrum [15]. When analysing voiced sounds, spectral tilt can be measured as the difference between the first and second harmonic amplitudes (i.e. H1-H2 / F_0-H2). Whereas the term spectral tilt is used to represent the slope of the spectrum, *spectral emphasis* is a measurement reflecting the relative mean energy values in the lower and upper halves of the frequency spectrum. These measurements may serve to show differences in voice quality [424, 417], accent [425] and emotion-related aspects [415, 426]. A study by Liscombe et al [281], for instance, suggested that spectral tilt (as well as type of phrase accent and boundary tone) may be useful in distinguishing between the qualitative valence (between positive or negative) of emotions.

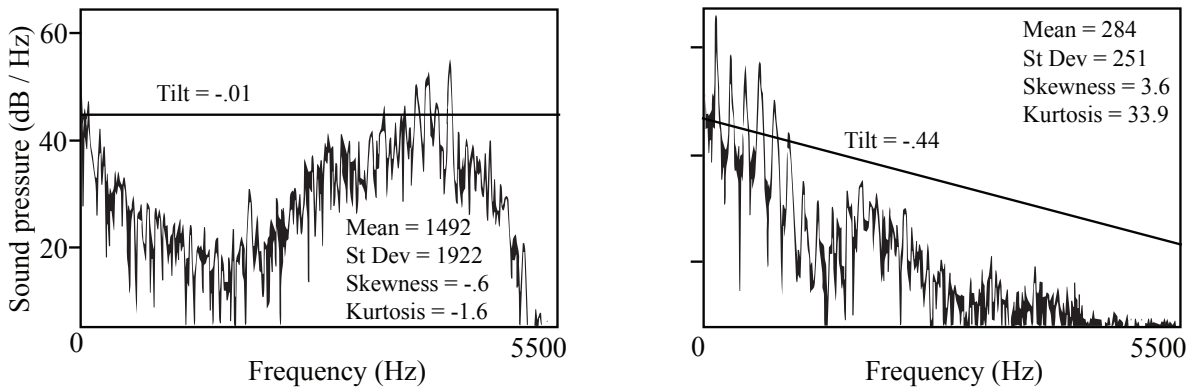


Figure 5.7: The Fourier spectrum with measurements for values of tilt, mean, standard deviation, skewness, and Kurtosis [15].

Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCC) [427] are standard features used in automatic speech recognition (ASR) designed to differentiate phones over quasi-stationary extracts of the speech signal [10]. The short-term statistics of MFCCs are commonly used to extract lexical information from speech, while long-term statistics are useful for paralinguistic analysis such as emotion [152]. Existing literature indicates that there are contradictory results. Steidl [10], for example, recommends that the use of MFCCs for emotion recognition can be useful as it has been successfully implemented [428, 429, 430, 431] to investigate *how* something is being said, rather than *what* is being said. Furthermore, he recommends that features with less reduction in earlier stages of the computation of MFCCs should consist of valuable information for emotion classification. Ververidis and Kotropoulos [241], on the other hand, observed a poor emotion classification performance [416, 432]. However, they suggest that this may be due to textual dependency, and the pitch-filtering algorithm used during cepstral analysis. Additionally, Yang and Lugger [296] stated that the emotional state of a speaker is unlikely to change as fast as phonemes. Their findings also suggested MFCCs to be less successful for emotion recognition (cf. [433, 432]).

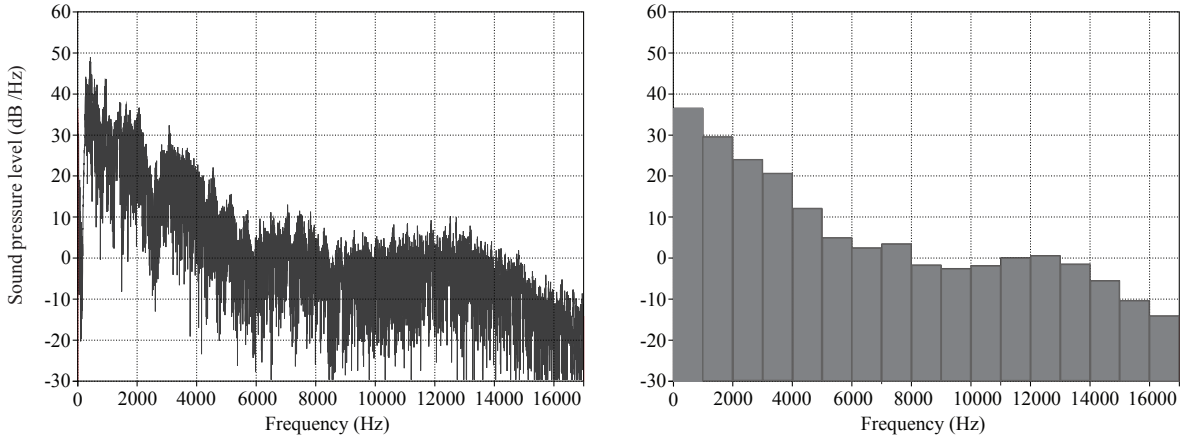


Figure 5.8: Long-term average spectra (LTAS) of speech utterance from female talker. The mean energy is shown across partitioned bands of width 1000Hz (left) obtained from the raw Fourier spectrum (left).

Long-Term Average Spectrum Analysis

As mentioned, spectral analysis can be conducted on a frame level (see Figure 5.5 (b)) or averaged over whole segments (see Figure 5.8). In *Long-Term Average Spectrum Analysis (LTAS)*, mean energy is usually extracted over segments of 30 seconds or more [42], which may make it useful for analysing emotional states that may lie over several speech segments. LTAS is generally quick and less susceptible to measurement error compared to most other measurements [139]. It cannot, however, pinpoint expressed emotion in speech segments [15] as it does not directly correspond to properties of sound at any exact moment [139]. Thus, it may overlook possible short-term acoustic cues that may be important when decoding emotional speech.

5.8 Automatic Emotion Recognition

With an appropriate speech corpus, work can be carried out to exclusively focus on recognising the state of the speaker by means of acoustic analysis (encoding studies). For automatic emotion recognition, a machine classifier is trained to classify emotion of a speech utterance based on the extracted acoustic features. Together with a label to represent the expressed emotion, the features are provided as input to machine learning algorithms [375] i.e. they are subsequently subjected to global statistical measures—global statistics are seemingly less sensitive

to linguistic information and are therefore useful for emotion in speech recognition [434]. In earlier studies, feature extractions consisted mostly of statistical measurements such as maximum, minimum, median, range and variability values on pitch, intensity, and duration-related features [435]. Alternatively, one can use mathematical transforms on a waveform, such as Linear Prediction Cepstral Coefficients (LPCC), Log Frequency Power Coefficients (LFPC), and Mel-frequency Cepstral Coefficients (MFCC). In recent advances, acoustic feature sets have become very large. For example, the feature set used for INTERSPEECH 2009 Emotion Challenge consisted of 384 features, the INTERSPEECH 2010 Paralinguistic Challenge included 1582 features, and the INTERSPEECH 2011 Speaker State challenge totalled 4368 features [178]. Although a large amount of features can be extracted, Ververidis and Kotropoulos [434] state that the performance of any classifier is reduced when all features are included.

5.9 Conclusion

As we set out to investigate how listeners perceive emotion from acoustic features, this chapter examined the most prevalent acoustical correlates of emotional speech found in the literature. Needless to say, a basic knowledge of speech acoustics is required to make informed choices about appropriate investigations on the perception of emotion in speech. The source-filter model of speech was first discussed (section 5.1) as is now widely used in emotion and speech research. Following this, the different acoustical aspects of emotion communication that have been shown to correlate to emotional expression were discussed, which can be subdivided into prosody (section 5.2), pitch (section F_0) (section 5.3), time (section 5.4), intensity (section 5.5), voice quality (section 5.6), and spectral features (section 5.7). Although this chapter provides general information for our research questions, it particularly contributes to exploring research questions four and five (RQ4 and RQ5).

6

Development of an Online Rating Tool

6.1 Introduction

In section 3.3.3, several existing tools used for the measurement of emotions were identified. Considerations in the development and delivery of these tools are essential for labelling emotional speech, one of the main goals of this thesis. It was pointed out in section 3.2.6 that labelling can be administered to a (usually small) group of expert judges or to a large group of undefined judges, also known as *crowdsourcing*. In the traditional sense, where expert raters are employed to rate speech, tools for the task are often delivered in-house. To harness the power of crowdsourcing, one generally outsources the task to a group of people via the web. With the assistance of web-based technologies, tools for rating emotional speech can be de-

veloped to deliver listening tasks to participants via web browsers. In the case where a small group of expert raters is used, the concept of what is being rated can be thoroughly revised prior to the task to ensure labelling consistency. This may not be practicable when targeting an undefined group of people online, however. Therefore, to increase the likelihood of labelling accuracy, much care must be taken with the participant's understanding of the concept that is to be rated, and the ease and straightforwardness of the task itself. For the development of an online tool, this chapter aims to address these issues, and will contribute to answering the following research questions:

RQ1: What are the practical prerequisites for carrying out large-scale listening tests?

The development of the tool considered several practical issues for labelling emotional speech on a large-scale basis. The tool is, after all, the backbone for effective labelling. To deliver a fully constructed naturalistic emotional speech corpus, in the next chapter (Chapter 7) it was set out to use the tool to obtain emotional labels from listening tasks. This task will contribute to answering research questions two and three (RQ2 and RQ3). In addition, it has been decided to use the tool as the basis for controlled experimentation to investigate nonverbal aspects of mood induced speech (Chapter 8), and thus contributing to answering research questions five and six (RQ4 and RQ5). Because the tool needs to be suitable for the experimental design, some alterations will be made. This chapter details the specifics for developing the tool for both tasks.

6.2 Descriptive Scheme

Having reviewed the different labelling schemes (see Chapter 3), it is evident that choosing an appropriate labelling scheme for annotating emotion in speech is not an obvious task, especially for natural spontaneous speech. At present, there is no consensual labelling methodology for describing emotions—the labels provided are linked with care to theoretical presuppositions that exist for emotion. In many cases, the labelling methodology a researcher settles on reflects the dataset description (e.g. eliciting type, delivered or observed emotions) and contingent

on the labelling task requirements, compromises methods applicable for either expert or non-expert participants.

The labelling scheme used in this chapter is the dimensional approach. This is for two main reasons. First, it is more suited for cross-studies in a wider context [251] because this approach avoids the complex issues often associated with subjective category labels for natural speech (see Chapter 3.4). Incidentally, the corpus is freely and publicly available to accommodate comparisons between studies. Second, the dimensional approach retains information about the correspondence and dissent of all ratings—rather than providing solely identical measures—which gives a more comprehensive knowledge on the level of agreement between ratings. The conclusions made in this thesis largely rely on these analyses.

The theoretical framework for the labelling scheme used in this thesis is in parallel with that used for the Feeltrace tool (see [9]), i.e. using the Activation and Evaluation dimensions. The method used here differentiates from the Feeltrace tool, as it does not require time-continuous evaluation, i.e. trace labelling (see also the work by [236]). This is because utterances of discrete periods of time (termed as quantised labelling [30]) of short length ($\sim$5 seconds) are used and it is assumed here that within the speech segment no changes in emotion occur, and are thus kept constant [436].

Given that a principal consideration in this investigation is the use of crowdsourcing (large-scale, non-expert listening groups), a labelling scheme that required comprehensive training was avoided. To assist the participant's understanding of each scale, the tool includes a detailed instructions page. A preliminary case study was carried out to ensure that the whole task was conceptually undemanding, and to verify the participant's understanding of each scale to be rated (section 6.3). Participants were presented with two *discretised* scales. A coarser granularity on each rating scale was opted for by discretising the dimensional scales into fewer categories—which also accommodates the requirements for machine learning. Each scale is divided into five colour-coded categories (see section 6.3, Figure 6.3), which runs from *Passive* to

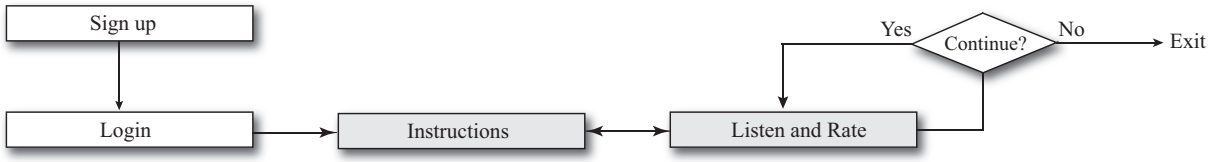


Figure 6.1: Flowchart of the web pages presented to the participant.

Active for the Activation scale and from *Negative* to *Positive* for the Evaluation scale. Previous studies have made use of categories such as 'very' negative and 'very' positive [9]; however, this study excludes these categories because of the nature of the speech material. Naturalistic speech conveys underlying emotion which can be considered milder and subtler expressions of emotion. Lastly, if a participant is unable to rate a clip, there is a "Do not rate" option, which, in itself, is documented as a rating.

6.3 Design and Implementation

First in this section, the initial design as it was administered for the first case study is described: labelling emotional speech (Chapter 7). Following this, a small case study that was carried out to ensure the efficiency of the tool is presented. Last, as the tool is later used for experimentation purposes (Chapter 8), the alterations that were made specific for the experiment design, and based on the feedback and experience from the case study, are described.

6.3.1 Framework for Case Study

As mentioned, the tool was delivered online in order to reach a large number of participants. There is an understanding among the research community that gathering ample data from the general population can be a demanding task. With this in mind, the tool has been developed and tested taking into account the end user's limitations (user-centred design (UCD)), ensuring ease-of-use and an adequate understanding for each scale to be rated. The participants' ability to use the tool with ease, and their understanding about the concept of each scale was given considerable importance as this would affect the quality of the labels. The design of the site (see Figure 6.1) ensures that the instructions were presented prior to the listening task, although

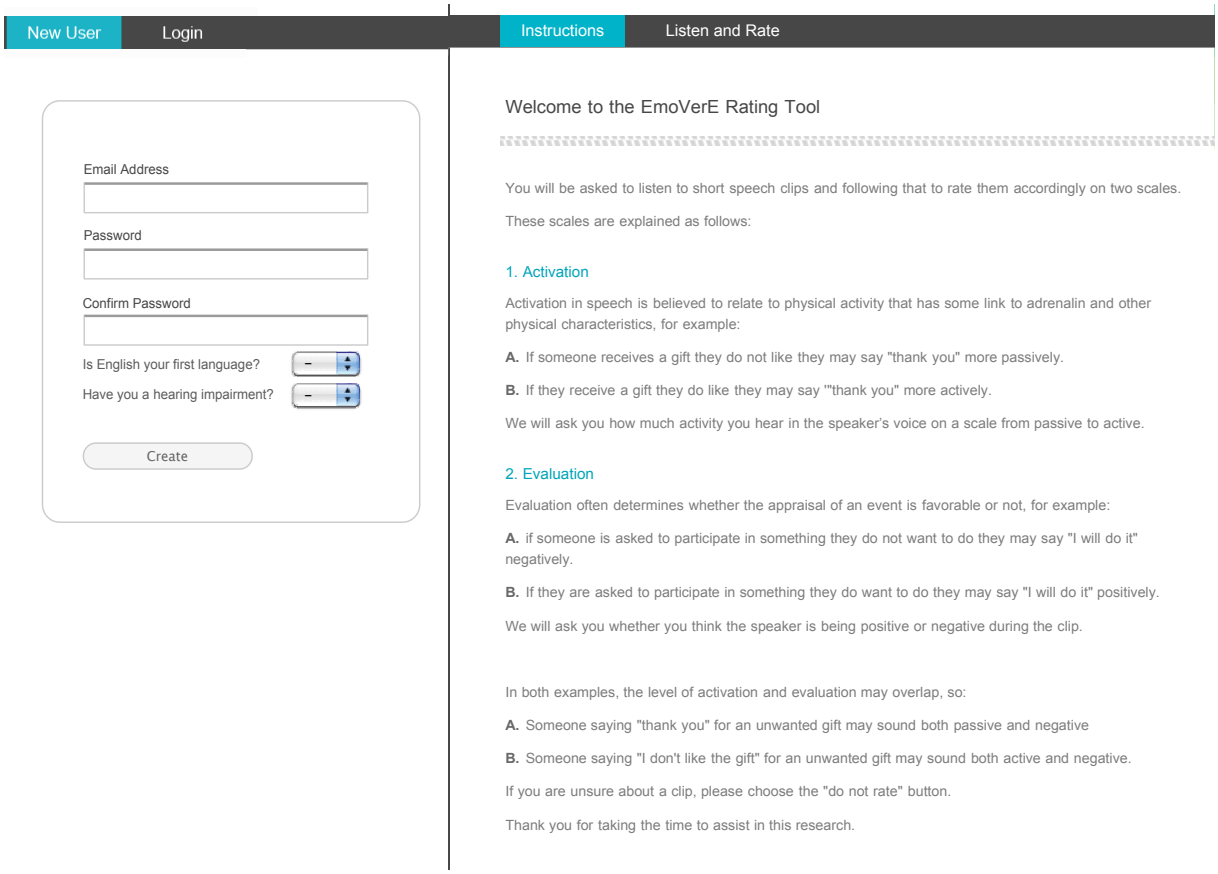


Figure 6.2: The sign up page (left) requires information on first language and hearing impairment. The instructions page (right) details the concept of Activation and Evaluation, with given examples.

the participants were able to refer back to the instructions at any stage during the task. Moreover, the tool was designed to be suited for repeated use to accumulate ratings on a continual basis. Furthermore, to encourage participation, minimum personal details were required.

For the initial design of the site used for the case study in Chapter 7, there were four main pages:

1. *Sign up:* participants were required to create a login account. To encourage participation, minimal details were required from the participants to take account for participant privacy concerns and to prevent participant impatience towards a daunting task. A sign-in ID (email address) and password was assigned, with mandatory information on first language and hearing impairment (on the left in Figure 6.2).

2. *Login:* participants logged in so that they could be identified with their previous tasks,

and, therefore, it kept track of speech clips that were already rated, ensuring each speech clip was only presented to them once.

3. *Instructions:* based on the dimensions of the circumplex model, Activation and Evaluation were presented on two scales. After login, the participant was presented with a description, and a written example conceptualising the Activation and Evaluation scales (on the right in Figure 6.2). The instructions, and the participant's understanding of them, were assessed with a preliminary survey (see section 6.3.2).

4. *Listen and rate:* the listening task was presented as three successive steps, i.e. listen to the speech clip and rate accordingly on both scales (see Figure 6.3). The scales were visually colour coded. Speech clips were randomised and taken from two subcategories: 'before' and 'after' segments of induced emotion (see section 7.2.1 for more details on subcategories). To ensure that the full audio clip had been listened to, the rating buttons were disabled until the audio player had reached the end of the speech clip. To rate a speech clip, both Activation level and Evaluation level had to be selected. If only one scale was selected the participant was notified and instructed to rate both scales. When the clip had been listened to, the participant could rate the current speech clip or choose not to rate it and continue to the next speech clip.

The number of speech clips presented in each session was kept to a minimum to prevent any fatigue and/or boredom effects, and, thus, decrease the likelihood of spurious ratings. For each session, participants were presented with six clips before given the option to exit. The number of clips presented was decided based on the feedback from the preliminary survey (section 6.3.2). Participants were asked to revisit and log back in at a later stage for continual rating. Participants were given the option to skip a speech clip if they felt they could not rate it by choosing "Do not rate" (DNR). If a participant chose DNR for three consecutive speech clips, they were notified and asked if they wanted to exit the session. A total of 160 speech clips were available for each participant to rate, and each clip could be replayed as many times as required by the participant. The participant's login details were stored in a MySQL database as it was felt it was easier to administer, and the ratings were gathered in an XML database as

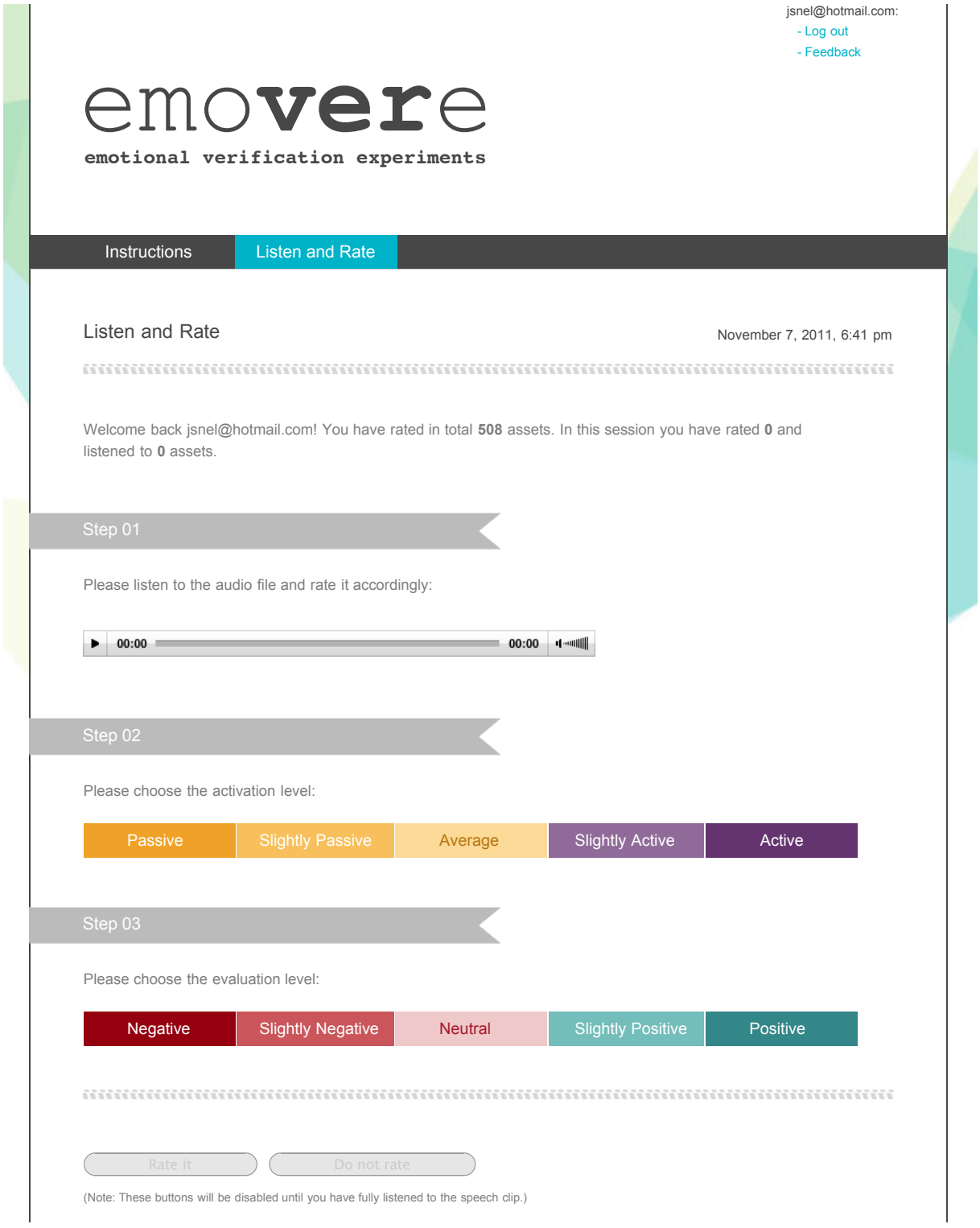


Figure 6.3: The main page of the web-based rating tool for rating speech clips. It includes an audio player, and colour coded scales for Activation and Evaluation.

this accommodated machine learning softwares. The data stored for each rating included the participant's email, the speech clip listened to and the associated ratings, and a timestamp of when the rating took place.

	Correct	Incorrect
Activation	6	1
Evaluation	6	1

(a) No. of correct and incorrect answers given for the multiple choice questions on the comprehension of the two concepts (Activation and Evaluation).

Demand	VL	L	N	H	VH
Mental	1	1	3	2	0
Temporal	0	3	4	0	0
Effort	3	1	2	1	0

(b) Subjective workload assessment, VL=Very low, L=Low, N=Normal, H=High, VH=Very high.

Table 6.1: Survey results

6.3.2 Design Validation

To validate the design of the tool prior to implementation, seven non-experts were surveyed, in the context of emotional judgement, to (1) assess their understanding of the instructions, (2) ensure they could setup up an account and complete the task without difficulties, and (3) assess the subjective workload when using the tool (see Appendix A for questionnaires). Participants were from a technical (college staff and other researchers) and non-technical (first-year journalism students) background. The steps taken for this were as follows:

1. Read instructions.
2. Answer questions about the definitions of both *Evaluation* and *Activation*—this is to validate the participant's comprehension of the two concepts.
3. Rate speech clips.
4. Complete assessment survey on workload.

For the activation question, six were correct and one incorrect. Similarly, for the Evaluation question six were correct and one incorrect (see Table 6.1a). It should be noted that the incorrect answers were from the same participant. This participant did not follow the order of the instructions in the above procedure. Instead, the participant read instructions, rated clips, and then answered the questions on Evaluation and Activation. Overall, it was concluded that there was a sufficient amount of understanding among the raters for the instructions of both scales.

After the rating task was completed, a survey (see Appendix A) based on the NASA TLX [437]—a subjective workload assessment tool—assessed the cognitive load on mental demands, temporal demands, and uncertainty, irritation, and stress (effort) while using the online rating tool. Overall, it was concluded that the cognitive demands were adequate (see Table 6.1b).

Participants were asked on the amount of clips that they would be willing to rate on a daily basis. Four participants chose three clips per day and three chose to increase the number. It was concluded that participants should be presented with three to seven clips at a time to prevent *boredom* and/or *fatigue* effects. Besides querying cognitive load, participants gave free-response feedback on any other information they felt gave difficulties. Accordingly, technical issues within reason—such as browser issues and password restrictions—were addressed.

Because there was the option of free-response feedback, a brief summary of some interesting remarks from the different participants are given below:

- Evaluation would be easier as binary.
- The definition of activation is easier to understand in terms of the dynamics of emotion.
- Scale for authenticity/genuineness could be introduced.
- There is a need for a baseline speech clip to compare against it.
- It was necessary to listen to some clips several times to hear the *tone of voice*, rather than the *linguistic content*.
- Others noted they assessed the clips along the scales according to the *linguistic content*.
- One participant expressed that the speech clips were “weird”.

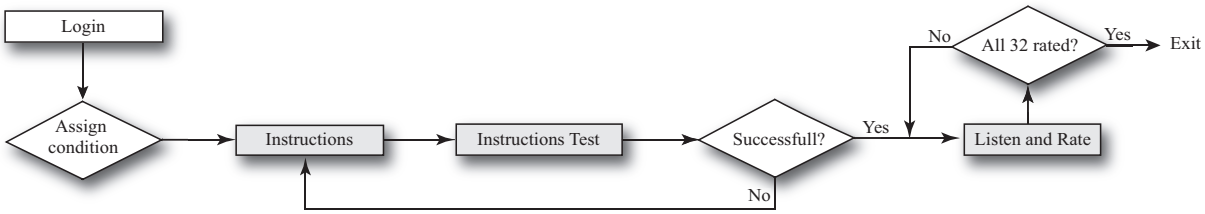


Figure 6.4: Flowchart of the web pages presented to the participant (experiment).

6.3.3 Adaptations for Experimentation

As mentioned, it was decided to use the framework of this tool for controlled experimentation to investigate recognition of emotion from nonverbal aspects. Some alterations were made based on the feedback and experience from the case study. More importantly, however, changes were made to the tool to be specific for the task and appropriate for the design of the experiment. The flowchart of the tool is shown in Figure 6.4. Again, a login page was necessary to track user ratings, and to ensure that the participants were assigned to the appropriate condition test. As illustrated, an additional subtask that followed the instructions page was included.

For the experiment, it may be necessary to take into account participant demographic information that could potentially be significant in the perception of nonverbal (prosodic) aspects of speech. Various studies indicate that individual differences in culture [321], age [438], gender [439], and laterality [440] can have an effect on the attentional bias towards either linguistic or prosodic content. For this reason, changes were made to the sign up/login page to include participant information on nationality, age, gender, and handedness. Moreover, the login page also required information on the participants group, as assigned by the researcher.

As already mentioned, speech was rated on two discretised 5-point (colour-coded) scales: Activation and Evaluation. For the participants to understand the concept of these scales, a page with detailed instructions on each scale was provided. Some participants who gave feedback during this case study (in Chapter 7) felt it was necessary to be presented with a baseline speech clip prior to the task to allow the participant to compare others against. The initial instructions provided for the case study did not include a viable baseline speech clip because no clips had

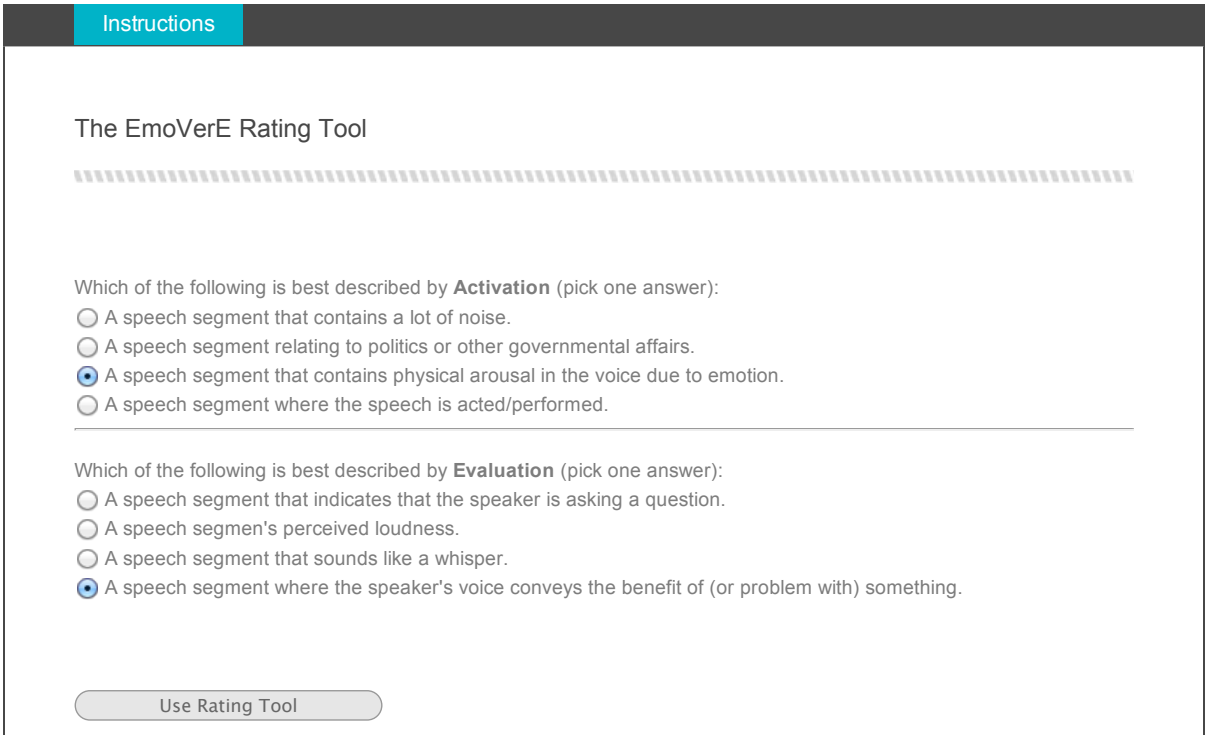


Figure 6.5: Instructions test page.

received any user ratings. Therefore, users could not be informed with a crucial label that was reliably applicable to the baseline clip. In the meantime, however, the ratings received from the case study provided us with a basis for a suitable label—the labels for all speech clips were determined by the ground truth values [441]. To allow the participant to be fully conversant with the task, an example speech clip with its corresponding value on the Activation and Evaluation scale (see Appendix B) was provided. The ground truth value of this clip was slightly active and slightly positive (3,3), and received 18 ratings in the case study[1]. The example clip was selected in compliance with the label's high inter-rater agreement[2]. This clip was not part of the stimuli rated in the experiment.

In the previous case study, 7 participants were surveyed prior to implementation to ensure an adequate understanding of the instructions was attained. Nevertheless, labelling accuracy was elaborated on by intervening the task with a 5-choice multiple questionnaire on the instruc-

[1]It should be noted that participants were informed that the provided example was based on the results from previous findings, and that the shown values did not necessarily indicate the correct chosen categories. Participants were informed of the values' subjective and inconclusive nature.

[2]The interval range was 0.89 on the Activation scale, and 0.78 on the Evaluation scale.

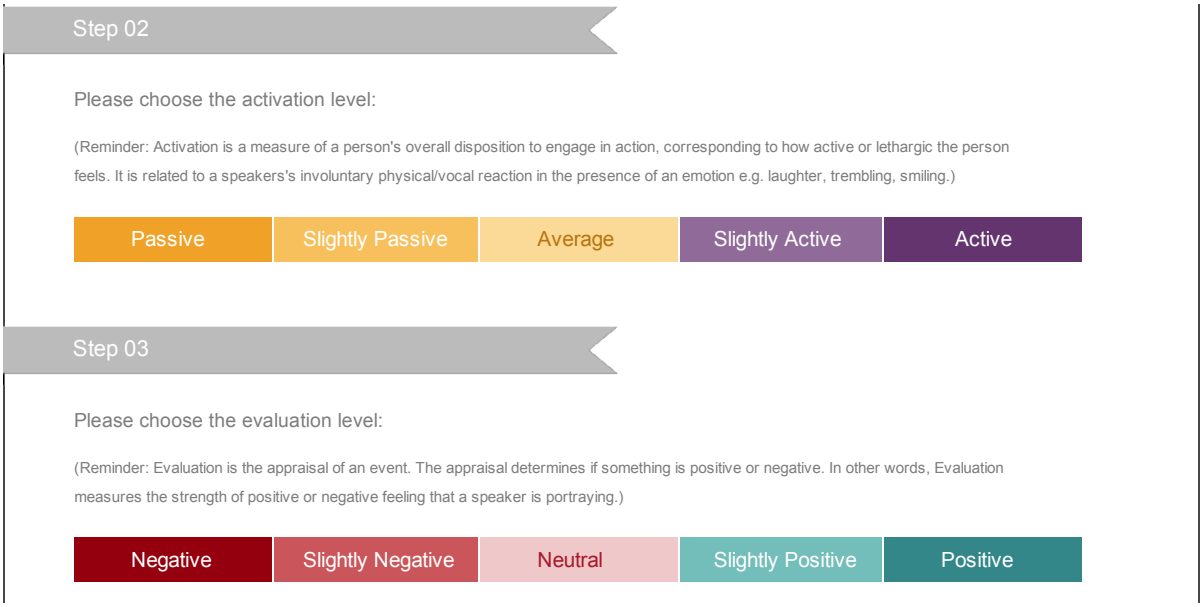


Figure 6.6: Listening task—with added notes on scale concepts

tions prior to the listening task (see Figure 6.5). Participants were required to complete the questionnaire correctly before continuing, and if the participant was unsuccessful for either of the scales, they were informed as a result and instructed to re-read the instructions carefully. Although this additional task may deter the participant from taking part after unsuccessful attempts, it nonetheless increases labelling accuracy. In retrospect, the task did not seem to prevent participants completing the task. Incidentally, all attempts were documented in a MySQL database—including the overall process of the task. Again from participant feedback, some indicated the need to regularly refer to the instructions page to recall the definitions for each scale. For this reason, the definitions were appended to the scale on the rating page (see Figure 6.6).

As participants were recruited, they were sequentially assigned to one of two groups, i.e. the first person would be assigned the task with the original/intact condition, the second to the filtered condition, and so forth. Participants could return to the task at any given time. When the user completed the first phase, the second phase was made active after two weeks had passed.

6.4 Conclusion

First, this chapter discussed the design of the online rating tool developed to be suitable for large-scale listening tasks, and thus contributing to answer research question one (RQ1). For this work, it was particularly necessary to make the tool accessible for laymen. For this reason, instructions were included. Prior to implementing the task, a short survey was carried out to ensure sufficient understanding. This chapter outlined the general framework as implemented for the case study in the next chapter and the subsequent experimental chapter. The tool is used in the case study to provide labels for a naturalistic, mood induced speech dataset. By providing the labels, it completes the construction of a naturalistic speech corpus (Chapter 7). This chapter also detailed the changes made to the tool in order to implement it for the experiment in Chapter 8.

7

Emotional Labelling: A Large-scale Perception Test

7.1 Introduction

One of the important observations from the previous chapters indicated that building suitable corpora for emotion in speech research remains a top priority. The build of an emotional speech corpora consists of two essential tasks: (1) collecting suitable and adequate emotional speech material and (2) assigning descriptive labels that correspond to the conveyed emotional content. Both procedures are effectively mutually dependent. That is to say, the emotional content in a given speech dataset needs to be established by carrying out listening tasks, yet at the

same time, the verification of its emotional content is largely dependent on the appropriateness of the labelling methods used. With a naturalistic speech dataset made available to us [442, 159], the first step is to provide well-founded user-verified labels that represent the perceived emotion for the individual speech clips. The aim is to achieve this by carrying out large-scale online listening tasks. For this task, participant magnitude and generality is a particularly important consideration, for which naïve listeners will be availed of, rather than solely 'expert' listeners familiar with emotion theoretical knowledge. The obtained ratings will establish the final labels—the quality being dependent on the ratings' reliability—from which the absence or presence of perceived emotional content can be determined, thus, exploring Mood Inducing Procedures' (MIPs) effectiveness as a means for eliciting spontaneous emotion. To this end, this case study will investigate the following proposed research question:

> **RQ2:** Can listeners adequately capture variation of activation and evaluation of emotion in naturalistic speech?
>
> **RQ3:** Can mood induction procedures provide naturalistic speech with sufficiently discernible levels of emotion?

This work aims to develop a naturalistic emotional speech corpus, with focus on audio quality and authenticity of emotional content. To our advantage, this work is founded on the knowledge of previous investigations [442, 159], and due to the challenges that the previous investigations met, i.e. the issues with obtaining a sufficient amount of occurrences for each clip, the rating strategy was re-evaluated (see section 7.2.1), which comprises some minor changes in the labelling methods and speech segmentation. In spite of the alterations, the speech material used for this study is extracted from the same source–speech recordings as an outcome of Mood Inducing Procedures [442]. The next section outlines the production of the speech stimuli used for this particular case study.

7.2 Methods

7.2.1 Stimuli Selection

In section 4.3 the type of corpora that are used for emotion in speech research were summarised. It was recognised that obtained spontaneous speech is in many cases recorded in unfavourable recording environments that can be problematic for effective acoustic analyses. The provided speech dataset was constructed with emphasis on elicitation of authentic emotions while focusing on the ideal recording environment, high audio quality being an aspect that is not often considered as a primary concern.

The Existing MIP Based Speech Dataset

Because there are inevitable restrictions in obtaining truly natural material and at the same time isolating the desired speech signal from unwanted noise, Mood Inducing Procedures (MIPs) provide for a convenient trade-off (see section 4.3.3). The speech material provided for this project is obtained from previous work [442, 159] that implemented MIPs. The procedures address the importance of audio quality and place great emphasis on the naturalness of emotional content. For its emotional content, MIPs were implemented that incorporated tasks with the need for achievement, and provided potential rewards for the participants—Success/Failure and Social Interaction MIP, and the Gift MIP. The construction of the emotional speech dataset considered several necessary aspects related to emotion induction, and among these were ethical attentiveness, authenticity of emotional content, and demand effects[1]. With regard to audio quality, the inducing methods were performed on participants positioned in soundproof isolation booths (see Figure 7.1) to minimise any unwanted acoustic factors. Speech was recorded at its highest audio quality of 192Khz/24-bit and for this investigation clips were downsampled to 44.1Khz. The equipment used for the MIP experiments comprised Neumann U87 microphones, Beyerdynamic DT 150 headphones, and a Pro Tools HD3 rig.

[1]Demand effects are those possibilities of the subject guessing the purpose of the procedure and hence act the desired emotion.

Figure 7.1: There are 4 booths available for recording, consisting of two different sizes. A recently installed sound proof booth is on the left, and the sound proof booth used for current speech material is on the right.

Subdivision of Dataset

The MIPs' efficacy of inducing non-neutral emotions can be determined by rating the dataset as a whole. However, this may not explicitly indicate that the MIP gave rise to these emotions. As an attempt to investigate the effectiveness of the MIP, the clips are split into two categories, *before* and *after*, according to the phase of the experiment from which it was extracted. One may assume a priori that over the duration of the experiment, occurrences of emotional episodes would increase as time unfolded. In other words, the greater the degree of participant engagement, the greater the degree of emotional involvement.

Vaughan explained in his thesis [159] that it is impracticable for the researcher to control for all emotion eliciting aspects over the course of carrying out MIPs. The researcher would have no control over any external engagement prior to the experiment that may have given rise to the participant's mood and/or emotional state. Inspecting the recorded material prior to rating and

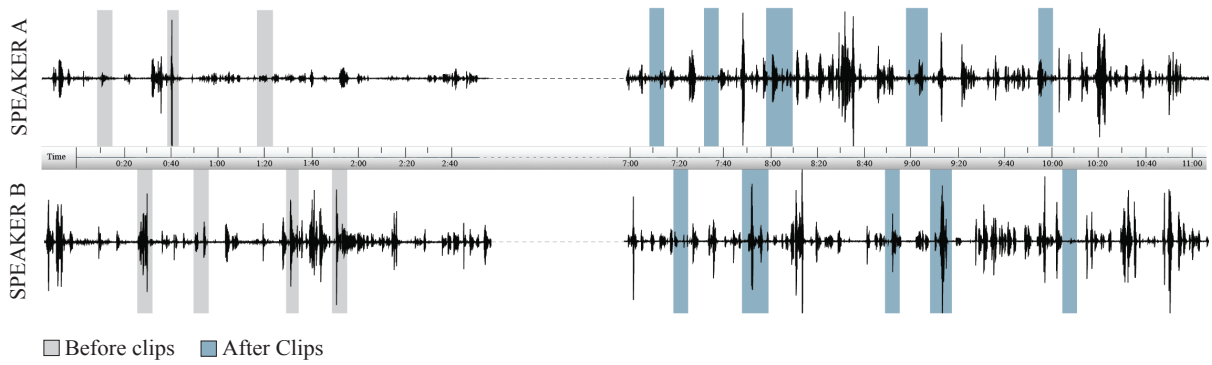


Figure 7.2: Extraction of speech clips categorised as *before* and *after*.

analysis, a subjective assessment suggested that before MIP manipulation commenced some participants immediately conveyed emotionally coloured speech, particularly with instances where the participants described and engaged themselves as friends. This is simply explained by social interaction (social interaction MIP). One may be interested in investigating the precise moments of elicited emotions as a result of the researcher's systematic manipulations. However, as Vaughan noted, the participant's response to any manipulation is not always predictable. Therefore, measuring the effectiveness of an MIP *manipulation* can be problematic or indeterminate, considering the difficulty of explaining every potential variable.

In spite of this fact, it is suggested to compare clips that are extracted from different stages of the experiment, and investigate how the MIP unfolded in time. Theoretically speaking, the participants should be less emotionally involved at the beginning stage, prior to experimental manipulations. To compare moments of the intended emotion elicitation, speech segments were extracted from different phases of the MIP experiment (see Figure 7.2).

Speech segments were extracted and assigned to one of two categories[2]: (1) those extracted from the beginning of the experiment, and, therefore, prior to/at the onset of the participant's immersion in gameplay—labelled as *before*—and (2) those extracted from the end to the middle of the experiment, when fully engaged in the MIP experiment—labelled as *after*. Subsequently,

[2]No linear or monotonic relationship was found between the individual clip's position in time and the observed rating. This would be expected because emotion fluctuates over time. It was more beneficial to compare two larger regions within the MIP experiment, i.e. *before* and *after*.

any differences in the rated content will be observed between the two categories and, therefore, investigate the effectiveness of the MIPs as an emotion eliciting exercise. At this point, it should be mentioned that the applied boundaries between 'before' and 'after' are vague, and the limitations imposed on interpreting the results are acknowledged. For the rating task, 80 speech clips were presented from each category, 'before' and 'after'.

Units of Analysis

Researchers need to consider how emotional episodes unfold and vary over time. This is mostly addressed with either labelling or segmentation in mind. In terms of labelling, it was mentioned that trace labelling is intended for this purpose (see Chapter 3.3.3). Alternatively, a single label is attached to a clip as a whole, with discrete time periods—termed as *quantised* labelling [30]. This approach is used in this thesis. The choice of units for segmenting speech, however, is a complex issue because it is difficult to determine where an emotional episode begins and where it ends. For now, speech clips are kept short to prevent transitions in emotional content in an utterance, and focus on singular states. In accordance with other studies [436, 158, 159], relatively short utterances were used, which ranged from 1 to 8 seconds in length (averaging $\sim$3 seconds). The segmentation process was carried out by the researcher based on intuitive assessment (see Appendix C for more details).

Rating Strategy

In order to conduct any constructive acoustic analyses, adequate amounts of labelled speech clips are necessary. However, obtaining large quantities of labelled speech has its practicable constraints. During his studies, Vaughan [159] identified the significance of choosing a suitable number of clips to rate. If too many clips are presented to the participant, the number of occurrences for a given clip will be too small for any statistically relevant analysis. In fact, this is coupled with the number of categories or discretised levels used. Vaughan reduced the number of clips for his second case study from 624 to 177. For this case study, 160 clips have been selected for rating and, to increase the likelihood of sufficient occurrences in the available

classes[3], the number of discretised levels used for each dimension were revised, using five discrete levels on each scale, instead of 21. Speech clips were extracted from eight different MIP sessions. The 160 speech clips (80 for each group) were produced by 16 different speakers (7m/9f).

7.2.2 Online Rating Tool

In the previous chapter, the development of the rating tool for this case study (Chapter 6) was discussed. To facilitate the use of crowdsourcing—and reach a large number of participants—the rating tool was delivered via the Internet. The objective of the tool was to have a simple but clean interface to make it easy for participants to understand and use. In order to achieve adequate labelling accuracy, the participants' ability to use the tool with ease, and their understanding about the concept of each scale was given considerable importance. The tool was designed with a simple but clean interface, and included a page with detailed instructions conceptualising each scale.

7.2.3 Selection of Subjects

It is argued here that the appointment of raters, and the accumulation process of ratings, is methodologically significant. Often in state-of-the-art research, rather small numbers of "expert" labellers are asked to participate in listening tasks. However, most research does not indicate explicitly what expertise the annotators have—annotators are usually researchers who are part of the wider field of emotional research. Cowie and Cornelius [16] argued that the wider, non-expert population can provide ratings that are equally valid to those of experts. Emotion is, after all, subjective in nature, and an important aspect of general communication between *all* humans.

The gathering of large numbers of annotators is rarely a primary research objective. As sug-

[3]There are five classes on both scales. On the Activation scale, classes range from Passive to Active, and on the Evaluation scale classes range from Negative to Positive.

gested by Tarasov et al. [160], the listening tasks for this study are outsourced to a large group of non-expert individuals, also known as *crowdsourcing* [161]. That is to say, the aim is to accumulate judgement ratings from a broader sample population that do not specifically require theoretical knowledge on emotion. For the online tool, the requirement of demographic information were kept to a minimum to encourage participation, and only sought information on English language as their mother tongue and hearing impairment.

Native versus Non-native Speakers

Because crowdsourcing is used to accumulate participants for this study, participants will be culturally diverse. It is widely known that there is cultural variation in emotional expressions [35, 43]. Studies have shown, however, that similar inference rules exist from vocal expressions across different cultures. A study by Scherer [43] showed that judges from nine different language speaking countries could infer four acted, categorical emotions from content-free speech, with a degree of accuracy substantially better than chance.

A study by Bryant and Barret [35] considered the idea that exposure to common emotion stimuli from, for example, mass media may be a factor in universal emotion inferences. For that reason, their study focused on participants from a South American indigenous population (the Shuar) and showed that those participants could identify emotions reliably from acted vocalisations conveyed by American native English speakers—the experiment implemented a task matching emotional spoken utterances to pictorial facial expressions.

In spite of the above-mentioned studies, culture-related variances in emotion-related studies are still being explored. For this study, it was deemed worthy to consider demographic information on native and non-native English speakers as it may have a substantial impact on the overall outcome, and, therefore, the resultant label. Besides, this study does not reliably compare to the aforementioned studies for two general reasons. Firstly, the material used in the two studies consisted of acted material, and as mentioned in section 4.3, it was argued that studies using

acted material must, to some degree, be differentiated from studies using spontaneous material, i.e. mood induced speech provided for this study. Secondly, there are numerous approaches to operationalising emotion and the studies above used categorical emotions, while this approach is that of the dimensional model (see Chapter 3 for information). Considering this, information based on English as a first language to operationalise native and non-native English speakers was gathered.

Hearing Impairment

Depending on the severity, it is clear that a participant's hearing impairment can have a strong impact on the outcome of each label. Very few participants, however, reported hearing impairments. Out of 107 participants who registered, four participants indicated they had some form of hearing impairment, three of whom only registered and did not rate any clips. These were therefore omitted. In the end, all four participants who indicated having a hearing impairment were excluded from the results.

7.3 Results

The accumulation process began in July 2011 and at the end of March 2012, 1243 pairs (Activation and Evaluation) of ratings received from 71 registered participants (averaging 7.77 ratings per clip) were reported [443]. The tool remained active to increase the potential number of ratings from those acquainted with the project after the publication was presented. In July 2012, there were 83 people registered[4] from whom ratings giving a total of 1707 ratings for both scales were received—excluding the 59 DNR ratings. The distribution of the overall ratings—before and after labelled clips—is shown in Figure 7.3 for the Activation (left) and Evaluation (right) scale. A total of 160 speech clips were rated, which gave us an average of 10.67 ratings per speech clip.

[4]In total, 107 people were registered but 24 of these did not rate any speech clips.

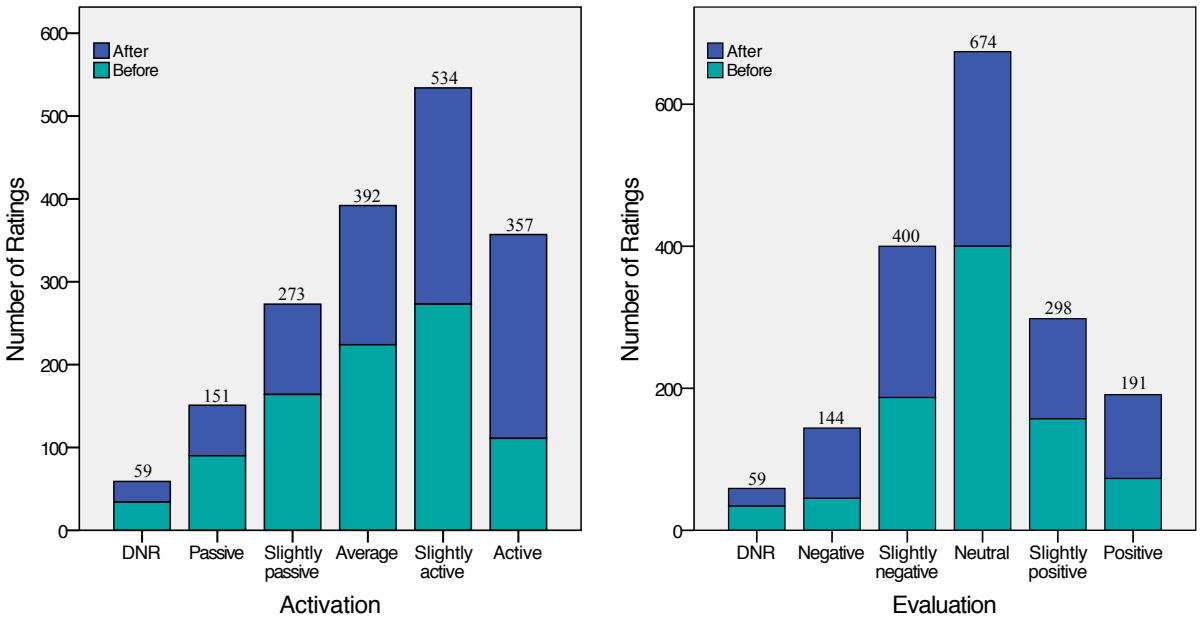


Figure 7.3: Distribution of the ratings received for the Activation (left) and Evaluation (right) scales, for clips labelled 'before' and 'after'. DNR = "Do Not Rate".

	Activation	Evaluation
Mean	2.39	2.00
Median	3.00	2.00
Mode	3	2
IQR	1.00	2.00
Overall SD	1.228	1.039

Table 7.1: Descriptive Statistics

For each scale, Table 7.1 shows the corresponding Mean, Median, Mode, IQR, and Standard Deviation (see Appendix D for further obtained values.). There is a slight negative skew on the distribution on the Activation scale, and the obtained mean value is 2.39 (SD=1.228). There is a slight positive skew on the Evaluation scale, and the mean value is 2.00 (SD=1.093). On the Activation scale, 392 out of 1707 were rated as 'Average' (23%) and 1315 as non-average (77%). On the Evaluation scale, 674 out of 1707 were rated as 'Neutral' (39%) and 1033 as non-neutral (61%). A Kolmogorov-Smirnov statistical analysis suggested a violation of the assumption of normality for both scales: Activation ($D(1707) = 0.212$, $p < 0.05$) and Evaluation ($D(1707) = 0.211$, $p < 0.05$) scales.

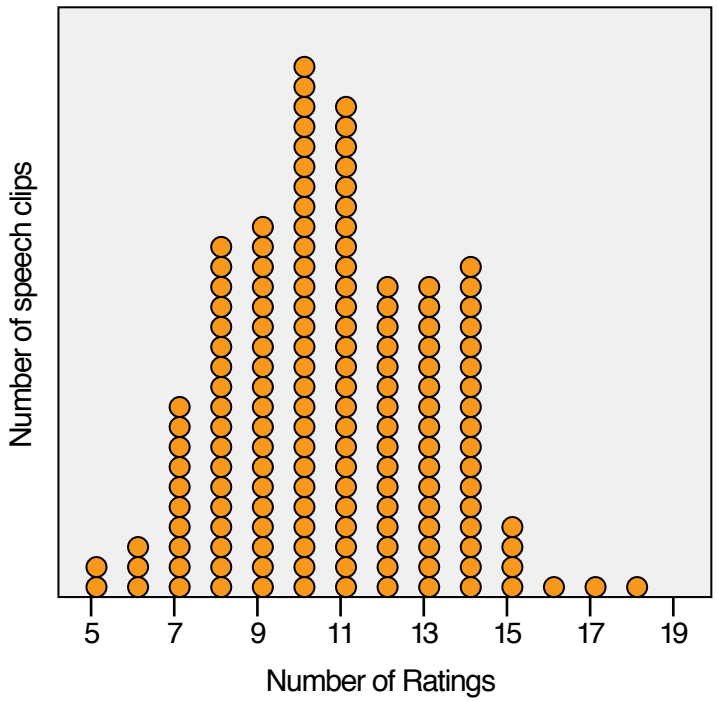


Figure 7.4: Dot plot for the number of ratings received.

DNR	No. of clips
3	1
2	8
1	40

Table 7.2: Number of DNR ratings received.

7.3.1 Rating Spread

As mentioned, participants were encouraged to rate all clips available but were advised not to overwork the task and return, if necessary, at a later date. Because there was no restriction on the number of clips that could be rated at any given time, the process needed to be randomised[5]. As would be expected, the received ratings were unevenly distributed. As already mentioned, a total of 160 speech clips were rated, which gave us an average of 10.6 ratings per speech clip. Figure 7.4 shows the spread of ratings received, where two clips received the minimum (5) and one clip the maximum number of ratings (18).

7.3.2 DNR Ratings

Participants were given the option to skip a speech clip if they felt they were not able to rate it by choosing "Do not rate" (DNR), which was also documented. Out of 1766 ratings, a total of 59 were DNR ratings (3%), which were spread over 49 speech clips. Table 7.2 illustrates

[5]Speech clips were randomised to minimise any order/practice effects

the overall breakdown which resulted in 1 clip receiving 3 DNRs, 8 clips receiving 2 DNRs, and 40 clips receiving 1 DNR. Because there is a relatively even spread of DNR ratings, and no specific clip received an exceptional number of DNRs, it was not considered significant for further investigation. In what follows in this Chapter, DNR ratings are regarded as 'missing' values.

7.3.3 Native versus Non-native Speakers

Of the 83 participants who took part, 38 were native English speaking (46%) and contributed to 1021 ratings, and 45 were non-native English speaking (54%) and contributed to 686 ratings. A Mann-Whitney (non-parametric)[6] test was performed to determine if there were differences in ratings obtained between native and non-native English speakers. Distributions of the ratings for native and non-native groups were similar on both scales, as assessed by visual inspection (see Appendix E). On the Activation scale, the test revealed no significant difference between the ratings obtained from native English speakers (Md = 3, n = 686) and from non-native English speakers (Md = 3, n = 1021), U = 347471, $z = -.282$, p = 0.778, r = 0.007. Similarly, on the Evaluation scale the test revealed no significant difference between the ratings from native (Md = 2, n = 686) and non-native English speakers (Md = 2, n = 1021), U = 340667.5, $z = -.997$, p = 0.319, r = 0.024. The following analyses report on the entirety of ratings, disregarding demographics of English as a first language.

7.3.4 Inter-rater Measures

Labels should only be assigned to speech clips that are validated in some form. The form of validation is subject to the description type considered by the investigator, i.e. cause-type or effect-type [9] (see section 3.2.5). The listening tasks performed for this study (effect-type) focus on the impression the speaker's way of conveying emotion has on the listener. In order

[6]The Mann-Whitney test is a non-parametric test. This was used because we are dealing with ordinal data. Many studies, however, report on the parametric equivalent. In most cases, a similar interpretation can be made.

to validate labels in effect-type (or judgement) studies, a high degree of agreement between all annotators needs to be associated with each label [240, pg. 175]. In this section, two aspects of inter-rater measurements are looked at. First, the overall consistency of order between ratings stretching over all clips is measured, using Krippendorff's Alpha (α), and second, the level of variation of the ratings for each individual clip by measuring standard deviations is determined.

Inter-rater Reliability

Krippendorff's α [17] for both scales, Activation and Evaluation, was computed. This is a reliability coefficient that measures how much raters agree on labels among multiple items—speech clips in this instance. It is most appropriate for the nature of the data for this thesis, as it is well suited for data that contains missing values (e.g. DNR ratings)[7] and for studies where multiple (instead of just two) raters are recruited, it can be used with any metric or level of measurement, and is applicable to either large or small sample sizes—it does not require a minimum sample size. Moreover, the single coefficient enables us to compare a variety of data with the same reliability standard. Krippendorf's α [17] is calculated as follows:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{7.1}$$

where D_o is the observed number of disagreements between raters:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck}\, \text{metric}\, \delta^2_{ck} \tag{7.2}$$

D_e is the expected level of disagreement by chance:

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_k\, \text{metric}\, \delta^2_{ck} \tag{7.3}$$

and o_{ck}, n_c, n_k, and n are arguments that represent the frequency of values in coincidence matrices.

[7] As seen in Figure 7.4. In this data, the number of ratings received are unevenly spread, which would be considered—by most inter-rater algorithms—as incomplete. Accordingly, Krippendorff's α accommodates for incomplete (missing) values.

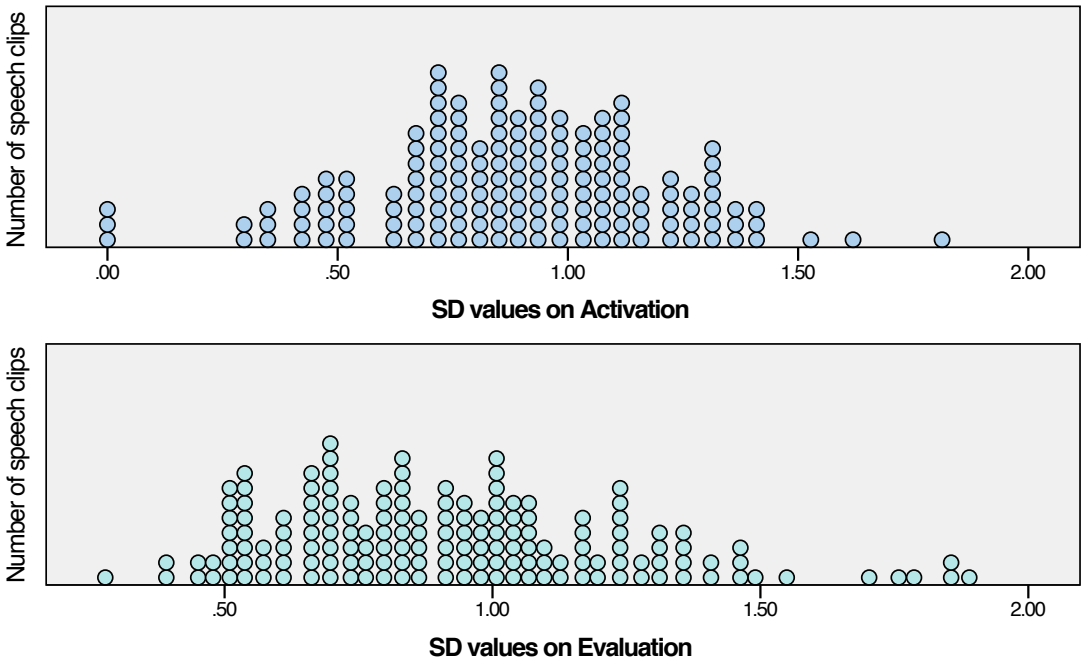


Figure 7.5: Dot plot for the SD values on each scale.

The observed value for α is interpreted as a level of agreement between 0 and 1, where 1 indicates perfect reliability and 0 the absence of reliability. For the 160 clips (or units), the maximum number of received ratings was 18, which in terms of Krippendorff's α relates to the number of observers—clips that received fewer ratings had missing values assigned to them. Table 7.3 shows the observed Krippendorff's α coefficients (ordinal) for each scale. The observed value on the Activation scale was higher than the Evaluation scale, with a score of 0.4268. On the Evaluation scale the observed α coefficient (ordinal) was 0.1999.

	Krippendorff's α
Activation	0.4268
Evaluation	0.1999

Table 7.3: Krippendorff's α [17] (ordinal) as a measure for inter-rater reliability for both scales.

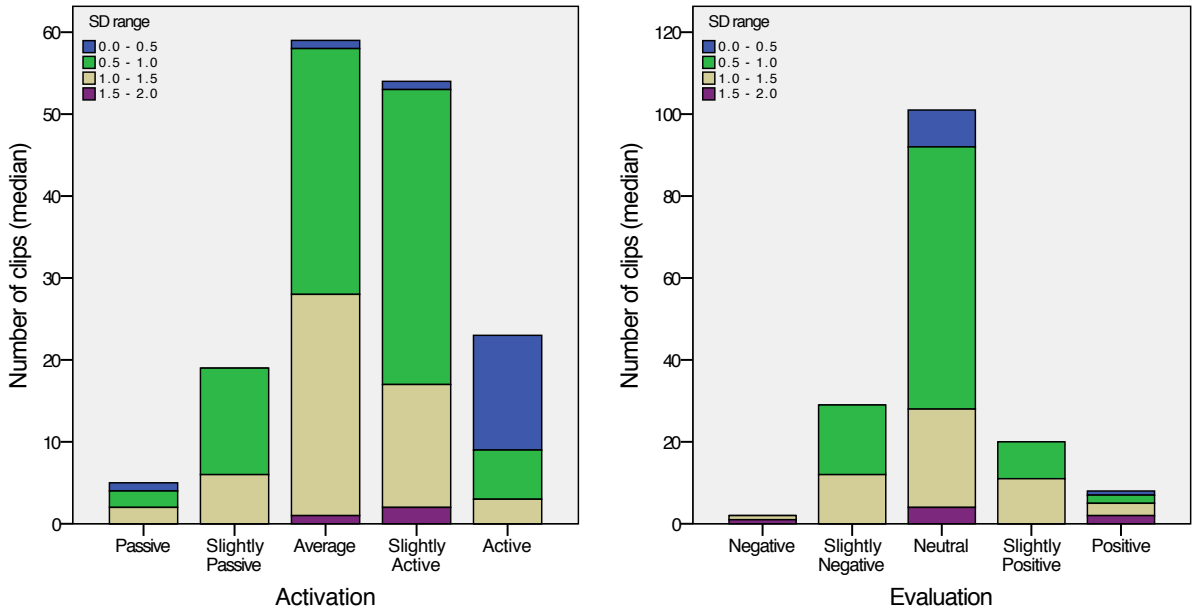


Figure 7.6: Distribution of clips with respect to the SD value and the median value obtained—the Activation (left) and Evaluation (right).

Inter-rater Agreement on Individual Clips

As mentioned in the previous section, Krippendorff's α is a measure of agreement determined from ratings received from multiple raters, across *multiple* clips. It does not allow for measuring agreement among raters for *individual* clips. Considering the reliability was low for both scales—in particular for the Evaluation scale—it would be advantageous to assess agreement levels on the individual clips. After all, some clips are more ambiguous and perhaps more difficult to rate than others, which is likely to have an impact on the overall reliability as measured above (section 7.3.4). To assess agreement levels on individual clips, the standard deviation (SD) of the ratings received for each respective clip was used. Figure 7.5 depicts how clips are distributed with respect to obtained SD values. The plots demonstrate that the majority of clips have an SD value between 0.5 and 1.00 for both scales (87 clips on the Activation scale and 92 clips on the Evaluation scale). The average SD value on the Activation scale is 0.893, and 0.9255 for the Evaluation scale.

Because the obtained SD values differ for all clips, it is worth examining the trend of SD spread more closely in relation to the emotional content of the clips. For instance, one can

Activation	Before	After	**Evaluation**	Before	After
Passive	90	61	Negative	45	99
Slightly Passive	164	109	Slightly Negative	187	213
Average	224	168	Neutral	400	274
Slightly Active	273	261	Slightly Positive	157	141
Active	111	246	Positive	73	118

Table 7.4: A comparison of ratings received for each class on each scale between the 'before' and 'after' labelled clips.

investigate in which rateable category the emotion identifier is most explicit, determined by the clips agreement level, which may provide some knowledge on the ambiguity of expressed emotion. One may expect that clips with no emotion—or rather 'average' on the Activation scale and 'neutral' on the Evaluation scale—may exhibit a trend of high agreement (low SD scores). Figure 7.6 demonstrates a cumulative bar chart of SD values within four different ranges with the respective median score. The Activation scale (left) shows that the highest agreements (SD range of 0.0 to 0.5) obtained were mostly for clips with a median value on the 'Active' class (coded as 4), and those with an SD range of 0.5 to 1.0 received a large number of clips for which the clips median value was 'Slightly Active' (coded as 3). On the Evaluation scale, the highest agreements (SD range of 0.0 to 0.5) were mainly observed for clips that were rated as Neutral, and similarly for clips with SD range of 0.5 to 1.0 (see Appendix F for histogram with mode values, and Tables for precise figures for respective median and mode values).

7.3.5 MIP Phase Comparison: Before and After

Furthermore, the difference between speech clips extracted from the beginning of the MIP experiment—and, therefore, before fully immersed in the experiment—and towards the end of the experiment were assessed. As mentioned above, the clips that were extracted from the beginning were labelled as 'before' and those extracted towards the end were labelled as 'after'. Table 7.4 compares the distribution of ratings for the *before* and *after* clips.

Differences between Individual Ratings

A Wilcoxon Signed Rank Test was performed to determine if there was a difference in ratings for the speech clips labelled as 'before' and 'after'. On the Activation scale, the test revealed a significant increase in the level of activation for the ratings obtained for the 'after' clips (Md = 3) compared to the 'before' clips (Md = 2), z = -6.951, p $<$0.001, with a small effect size (r = 0.168). However, on the Evaluation scale the test revealed no significant difference between the ratings obtained for the 'before' clips (Md = 2) and 'after' clips (Md = 2), z = -1.534, p = 0.125, r = 0.037.

Differences between Inter-rater Reliability

Table 7.5 shows the obtained α measures for clips taken from the start and end of the MIP experiment. It shows that the agreement coefficient is higher for the 'after' clips on the Activation scale, but slightly lower for the 'after' clips on the Evaluation scale.

	Before (α)	After (α)
Activation	0.2978	0.51
Evaluation	0.2072	0.1907

Table 7.5: Krippendorff's α [17] (ordinal) values for experiment phase (before and after) on both scales.

Differences between SD Values

SD values between the clips from both conditions—labelled as 'before' and 'after'—were also compared. In this case, we are dealing with a continuous dependent variable (SD). A preliminary test was carried out to assess violations of normality. (Preliminary analysis for each scale ensured no violation of the assumptions of normality.) Each scale contained one outlier, but by inspection neither revealed an extreme value, and were therefore included in the analysis. The assumption of normality was not violated (differences between the SD scores of the 'before' and 'after' clips were normally distributed), as assessed by the Shapiro-Wilks test for the acti-

vation (p = 0.490) and evaluation (p = 0.424).

Accordingly, a paired-sample t-tests was conducted to evaluate whether there was a statistically significant mean difference between the agreement level (as measured by SD) of the 'before' and 'after' clips. For the Activation scale, the test indicated a significant *decrease* in SD measurements (contrastingly interpreted as an increase in agreement) between the 'before' clips (M = 0.9671, SD = 0.253) and the 'after' clips (M = 0.8188, SD = 337), t(79) = 3.142, p = $<$0.002 (two-tailed), d = 0.3512. The mean decrease in SD measurement was 0.148 with a 95% confidence interval ranging from 0.05436 to 0.24234. The eta squared statistic (.111) indicated a large effect size.

For the Evaluation scale, however, the t-test indicated a significant *increase* in SD measurements between the 'before' clips (M = 8267, SD = 0.24482) and the 'after' clips (M = 1.0242, SD = 0.362), t(79) = 3.142, p = $<$0.0005 (two-tailed), d = 0.4648. The mean increase in SD measurement (interpreted as a decrease in agreement) was 0.198 with a 95% confidence interval ranging from $-.29205$ to $-.10297$. The eta squared statistic (.179) indicated a large effect size.

7.3.6 Speech Clip Duration and Agreement Level

As mentioned above (section 7.2.1), the speech clips used for this case study are of relatively short lengths. One may suspect that the shorter the speech clips are the more difficult they are to rate, which may therefore result in lesser agreement (or more DNR ratings.) For this reason, the relationship between SD measures and the clip duration will also be investigated.

A Spearman's rank-order correlation (non-parametric) was performed to assess the relationship between the SD of the ratings on a clip and the duration of the clip. The results indicated that the correlation between SD and clip duration was not statistically significant for the Activation

scale ($r = 0.080$, $p = 0.314$) nor the Evaluation scale ($r = 0.011$, $p = 0.884$).

7.4 Discussion

In this chapter, the aim was to obtain ratings from a large group of non-expert individuals, yet this proved to be time consuming—ratings were collected over a period of a year, gradually declining in number from the beginning. A total of 83 participants contributed to a total of 1766 ratings, 59 of which were "Do not rate" instances (3%). Although participants were asked to return regularly and rate all 160 clips over a given period, the majority rated $<$20 speech clips, the average being 25. However, using fewer (5 discretised) categories on both scales, and not availing of a surplus of rateable speech clips, gave a satisfactory spread of ratings over all clips coupled with sufficient occurrences for each class on each scale. This provided us with adequate substance with which to analyse the findings.

7.4.1 Distribution of Ratings

For each scale, Figure 7.3 shows the spread of ratings over all classes. The evaluation scale (right) contains a large number of neutral ratings, which would be somewhat expected (see also [444]), gradually decreasing towards positive and negative classes. As mentioned earlier, 39% of the ratings were rated as Neutral and 61% as non-Neutral. On the Activation scale (left), however, the majority of ratings appear in the Active classes. The large number of active, non-neutral clips could be explained by the nature of the tasks carried out during the MIPs. Overall, the histogram demonstrates the presence of emotional content in a sufficient number of speech clips. One could suggest therefrom that MIPs were, for this speech dataset, successful in eliciting emotion. However, everyday speech is generally coloured with emotion expression, so it may not explicitly indicate that the MIPs gave rise to these emotions.

7.4.2 MIP Phase Comparison

As mentioned in section 7.2.1, the investigation was broadened by assigning the clips to one of two categories according to which phase of the experiment it was extracted. The investigation revealed a significant increase in the level of activation for the ratings obtained for the 'after' clips, but showed no significant difference for the evaluation scale. This would suggest that as the experiment unfolded, the more the participant became emotionally involved, or at least in terms of activation.

Of the speech clips extracted from the 'before' and 'after phase, both Krippendorff's reliability and SD mean differences gave similar interpretations. The SD measures on the Activation scale indicated that there was a significant increase in agreement for those clips extracted towards the end of the experiment (after clips). In other words, the more emotionally involved the participants became—or the more active speech became—the less difficult they became to identify. Upon evaluation, however, there was a significant decrease in agreement for those clips extracted towards the end of the experiment, indicating that as an emotional expression became more extreme, the more ambiguous or less uniform ratings were. Perhaps, one may expect that neutral speech would be easier to evaluate than negative or positive speech because emotional speech can be ambiguous.

7.4.3 Inter-rater Measures

Inter-rater reliability for the task was rather low for both scales, where α was 0.4268 for the Activation scale and 0.1999 for the Evaluation scale. These scores appear to be in line with other studies that report on α measures lower than 0.55 [445, 310, 446]. In contrast to the findings here, where the reliability coefficient was lower for Activation compared to Evaluation, the study by Truong et al. [445][8] found that agreement was mostly higher for the valence scale,

[8]Their study compared audio and visual channels and obtained overall agreement levels ranging from 0.12 to 0.48

and reported agreement levels on their speech recordings of around 0.12 for Arousal (active vs. passive) and 0.32 for Valence (positive vs. negative). The low inter-rater reliability measures were expected, and in accordance with what Craggs and Wood [310] acknowledged, the results suggest that identifying emotion reliably is a difficult task, or at least that emotion is not necessarily perceived uniformly among individuals. Rating *spontaneous* speech is a demanding task because the emotions are mostly underlying—milder and subtler than full-blown prototypical expressions. It is, after all, difficult to obtain natural speech with intense states through MIP experiments because of the restrictions on ethical matters. Moreover, the difficulty of the task may be, to some extent, related to the use of short speech clips ($\sim$5 seconds). As a reminder, short clips were used to minimise emotional transitions and overlapping emotional states. Although research has suggested that participants can effectively recognise emotions as short as $\sim$5 seconds [159, 436], the potential effect of speech clip duration will also be briefly discussed. First, as already mentioned no particular clip received an exceptional number of DNR ratings. In fact, the proportion of DNR ratings was only 3%, which would suggest that participants had few difficulties rating the recordings. Second, comparing the relationship between SD measures and speech clip duration indicated that there was no statistically significant correlation, suggesting that clip duration—1 to 8 seconds—did not influence agreement levels (SD), which can be interpreted as the level of ambiguity or difficulty associated with the clip's identification task.

The standard deviation measure is often used as a measure of agreement [152, 32]. A major advantage of this measure is that it allows us to analyse agreement levels on individual clips. The average SD value on the Activation scale was 0.893, and on the Evaluation scale was 0.9255. Unlike the Krippendorff computation for reliability, the average SD values demonstrate a small difference between the two scales. The average SD for the VAM corpus, which also used five discretised classes, was reported at 0.29 for valence, and 0.34 for activation [152]. Again, in contrary to the findings here, their results showed that evaluation received a smaller SD measurement—showing better agreement. It is difficult to explain why the average SD results here are higher than that of the VAM corpus. The most credible explanation would probably be

the type of emotional content present in the speech material—similar rating schemes, and similar sentence durations (averaging 3 seconds) were used. First, a subjective assessment made by the researcher on the two types of speech data suggests that the emotions portrayed in this speech data are subtler compared to those of the VAM corpus. Secondly, in terms of linguistic content, the corpus used in this study seems to provide little information to indicate emotional content. Thirdly, comparing to the results here, the VAM corpus contains quite a large number of negative ratings, which may suggest that negative speech is easier to identify.

Figure 7.5 showed how clips were distributed with respect to obtained SD values. It demonstrated that on the Activation scale, in proportion to the number of ratings received for the corresponding class, the highest agreement was among clips with median value located in the Active class, and on the Evaluation scale highest agreements were among clips with median value located in the Neutral class[9]. On the Evaluation scale, the extreme classes—Negative and Positive—received the lowest agreement. This suggests that active speech is identified more reliably but appears more ambiguous on the Evaluation scale the more extreme the observed classes are. A study by Schröder [287] investigated how written text and synthesised speech were perceived on the activation and evaluation dimensions. The study reported that through text on average SD measures were higher for Activation, i.e. Activation was more difficult to judge in contrast with Evaluation. In terms of prosodic cues, Activation was portrayed most successfully while evaluation proved to be more difficult. Schröder acknowledged that different aspects of emotion are portrayed through different channels—acoustic and linguistic. It seems that linguistic content is a strong contributor in emotion identification.

7.5 Conclusion

The purpose of this chapter was to investigate the existence of perceived emotion in an MIP based speech dataset, by performing listening tests on a large-scale basis. It aims to answer

[9]Mode measures gave similar results. See Appendix F for distribution of SD value respective to the Mode value

questions two and three (RQ2 and RQ3). For this case study, an online rating tool was developed to accumulate ratings on a large-scale basis and to ensure that the participant's needs and limitations were met in order to reinforce rating accuracy (Chapter 6). As mentioned in the last chapter, in order to successfully deliver the tool, it was important to ensure that naïve listeners could adequately understand the concept of dimensional rating. The feedback obtained prior to implementing the listening task (see section 6.3.2) suggested that participants could adequately rate on dimensional scales. To some degree, this was supported by the results obtained in this case study, given the low number of DNR ratings obtained (section 7.3.2). The spread of ratings indicated (see 7.3.1) that participants could perceive emotion variation on Activation and Evaluation dimensions. Based on inter-rater agreement measures (section 7.3.4), however, it appears that the Activation scale is more reliable. Nevertheless, it is difficult to determine if this is due to the validity of the descriptive scheme, or due to the content of the speech data itself, where emotion portrayals are of a subtle kind. The mutual dependency in verifying the emotional content in the given speech data and assuring the appropriateness of the labelling scheme can be somewhat seen as a paradox. Investigating the dataset as a whole and then subdivided—based on a priori assumptions—gave us a foundation to examine the MIPs' effectiveness, while simultaneously demonstrating the appropriateness of the labelling scheme. Overall, it was concluded that the obtained results support both the MIPs practical validity, and support the chosen labelling scheme. In line with other findings in the literature (section 7.4.3), it is recognised here, however, that identifying emotion is fundamentally a difficult task, especially for spontaneous emotion of a subtle nature. The overall level of agreement, as tested by Krippendorff's reliability coefficient and SD measures, is below ideal, which appears to be more evident on the Evaluation scale. It was mentioned in the last section that Evaluation is more dependent on the linguistic channel, which could be a contributing factor to its low agreement levels.

The work in this chapter was part of an ongoing corpus-building project. The outcome provided ratings linked with the provided MIP based speech material. The work undertaken supplied a

deliverable naturalistic emotional speech corpus[10] made freely available to the general research community. In conclusion, this chapter contributed to answering research question two and three (RQ2 and RQ3).

[10]The corpus includes the DIT Mutual non-disclosure agreement form (to be signed), speech files of 44100 Hz sample rate, obtained ratings for each speech file and each scale (including user IDs), and Ground Truth labels for Activation and Evaluation.

8

Judging Emotion from Nonverbal Aspects of Naturalistic Speech

8.1 Introduction

The case study carried out in the last chapter demonstrated that there was a sufficient amount of perceived emotional content in the existing speech dataset and, therefore, it was concluded that the MIP procedures were successful in inducing non-neutral emotional states. The accumulated ratings determined labels for each speech clip, and, as a result, a structured and functional emotional speech corpus was built. At this stage, one could investigate which acoustic parameters correlate with perceived expressed emotions (encoding study), based on the idea that

these acoustic correlates convey expression independent of the linguistic information. Indeed, there are ample influential studies that suggest probable acoustic parameter sets. To date, however, research has not yet established a reliable correlation between the acoustic signal and the expressed emotion. And, as discussed in section 4.4, this may be due to the incorrect or incomplete selection of acoustic parameters. Alternatively, as Mark Tatham and Katherine Morton [378] suggest, it may be due to the fact that listeners do not perceive emotion solely from the acoustic signal. Being in agreement with the latter, this concept will be explored in this chapter. To begin with, perhaps a basic distinction should be made between the *linguistic* and *acoustic* aspects of speech. In effect, this study is concerned with *what* someone is saying and *how* it is being said. Both acoustic and linguistic aspects of speech, however, are intrinsically fused, which makes empirical evaluation of one aspect in complete absence of the other more or less methodologically impossible. Despite these constraints, this study attempts to isolate the acoustic aspects of speech, while minimising any loss of salient acoustical features known to convey affect.

To enable applications to recognise and extract information on emotion from a speech signal, it is imperative to have an adequate understanding of human perception of emotional speech. To investigate speech signals, it is difficult—or in most cases impossible—to solely remove or isolate the fundamental psychoacoustic elements (such as pitch, loudness, timbre) to investigate its role because the associated physical properties (e.g. waveform, spectrum, intensity level, frequency, etc.) are potentially inclusive. To make progress in understanding the vocal communication of emotion, it is necessary to investigate voice cues as contributing elements part of the structured whole, rather than independent elementary components. The concept of perceptual organisation may be better illustrated visually (see Figure 8.1) in a manner analogous to the *Gestalt* principles [447]. Gestalt principles are often explained with the phrase “unified whole”, which states that complex systems are inherently irreducible and cannot be explained by their component parts alone. Instead, a holistic approach is needed. For a hypothetical example, Figure 8.1 illustrates different coloured circles that represent a set of acoustic features (pitch, speech rate, intensity, and spectral). The visual representation of the white box in Figure

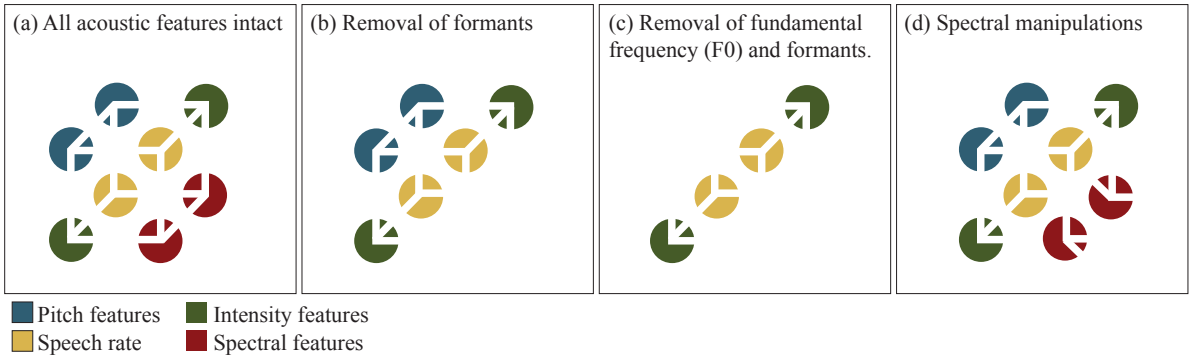


Figure 8.1: Feature sets analogous to Gestalt principles

8.1 (a) symbolises the emotion present in the speech signal, and as illustrated, its appearance is somewhat subjective and ambiguous, although discernible. By manipulating the acoustic signal, one may remove spectral features such as speech formants, yet retain prosodic features such as the pitch, intensity, and temporal related features (Figure 8.1 (b)), leaving a sufficient amount of information to perceive the emotion. However, if removing salient properties simultaneously such as pitch and certain spectral features (Figure 8.1 (c)), the conveyed emotion will be imperceptible. Despite these constraints, it is possible to systematically manipulate certain features, such as removing some of the spectral content by low-pass filtering the speech signal (Figure 8.1 (d)), quantifying the relevant effects on perception, and comparing its strength to other features. The main objective of the experiment presented in this chapter is to investigate how emotion is perceived in nonverbal aspects of naturalistic speech. In order to mask the verbal content, certain acoustic features need to be removed or manipulated. In this regard, the aforementioned example is taken into consideration, as some paralinguistic features will inevitably be distorted or degraded. Because these procedures affect different acoustic cues, the role of certain voice cues can be investigated in this vein.

8.2 Overview

The objective of this experiment is to segregate the acoustic channel and investigate its relative contribution in conveying emotion. In this regard, how subjects infer intact speech against

altered speech, i.e. intelligible speech (acoustic and linguistic cues), is compared against unintelligible speech (acoustic cues only). To attain unintelligible speech, the linguistic content is masked by applying a low-pass filter on the speech signal. This technique will allow us to explore the following research questions:

RQ4: Does nonverbal naturalistic speech convey Activity and Evaluation levels that are recognisable to listeners?

RQ5: How do ratings from two perceptually different conditions (verbal and nonverbal speech) compare?

A general theme of this thesis is the acoustic correlates of perceived emotion. It has long been recognised that acoustical patterns, such as intonation, rhythm, and vocal intensity, signify paralinguistic cues that have communicative functions to express a person's emotional state. Numerous acoustic features correlate with emotional speech, but the extent to which each feature influences perception of emotion in natural, spontaneous speech—or more specifically mood induced speech—is still being investigated. It remains uncertain about which of the two aspects, acoustic or linguistic, is more significant in expressing emotion, what each aspect's relative contribution is, and whether each communicative function corresponds invariably with the other. This chapter seeks to explore some of these issues, which effectively questions the validity of correlating acoustic parameters with labels that are representative of both acoustic and linguistic cues.

8.3 Methods

For the main experiment, participants were asked to take part in two separate listening tasks. The two tasks were performed two weeks apart. For one task, they were asked to rate speech in its recorded form, i.e. intact (non-filtered) and comprehensible, while for the other task they were asked to rate speech that was manipulated (low-pass filtered) to make it incomprehensible. Before full implementation, the task was carried out initially with four people to monitor

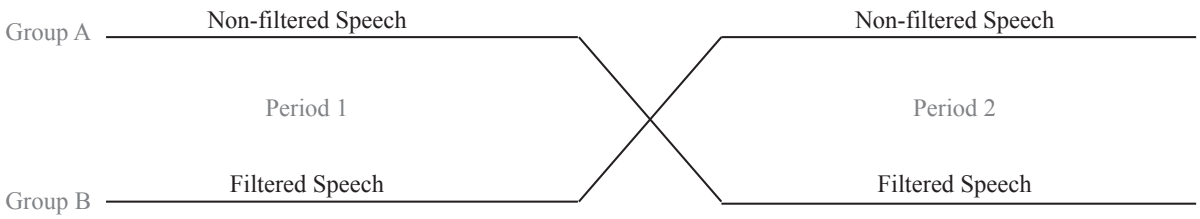


Figure 8.2: Experimental design: Crossover study.

the process to ensure the participant's full understanding and to confirm that all data was accumulated correctly.

8.3.1 Design and Implementation

The experiment consisted of a within-subject (repeated-measures), 2-period crossover design (see Figure 8.2). A crossover study is a longitudinal study that reduces confounding covariates—each participant serves as their own control—such as order effects, and individual subject differences, thus enhancing statistical power. To achieve this, subjects were randomly assigned to one of two groups. The groups were presented the stimuli conditions in different orders. The first group rated the non-filtered speech on the initial task and the filtered speech on the alternate task, and vice versa for the second group. The tasks were administered two weeks apart to reduce the subjects' retention of the speech tone (or F_0) from the stimuli in the first task (priming and/or carry over effects). For each task, each participant was asked to rate 32 speech clips in a phase. Therefore, for each participant, a total of 64 speech clips were rated if both phases were completed. Speech clips were randomised to avoid stimuli order effects. As before, participants were given the option to skip a speech clip if they felt they could not rate it by choosing "Do not rate", but each clip could be replayed as many times as the participant wanted. It was expected that the task of rating the filter condition would be somewhat more difficult as the intact speech. The speech stimuli were presented to the listener via the web-based rating tool as described in the previous case study (see Chapter 7) with modifications made to the tool to serve the experiment design. Participants were instructed to do the task using headphones in a quiet location to keep extraneous noise to a minimum. To emphasise the importance of performing in a quiet location, participants were asked to switch off any enter-

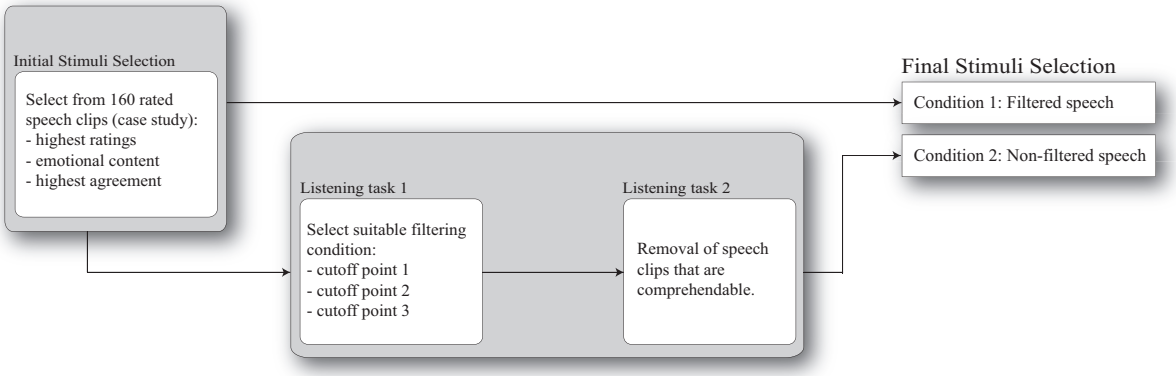


Figure 8.3: Work flow for stimuli selection.

tainment devices such as TVs and radios, and to minimise any other disturbances while doing the task.

8.3.2 Stimuli Selection

The stimuli used for this experiment are derived from the same dataset that was used for the previous case study (Chapter 7). The ratings obtained from the case study will specify the selection process of the stimuli for this experiment. To draw up the required stimuli for the two conditions (original and low-pass filtered), 3 preliminary tasks were carried out:

1. Selection of suitable intact (original) speech stimuli
2. Evaluation of an effective filter condition to administer
3. Verification of incomprehensible spoken dialogue in the selected filtering condition.

It should be noted that the conclusions drawn from the experiment are not contingent on the clip selection strategy. Instead, the selection preference acts as an accompaniment to the proposed research questions. The three stages (see Figure 8.3 for the workflow) that determined the stimuli for this experiment include a preliminary analysis for the selection process based on the previous obtained ratings, and two subsequent listening tasks, using two independent groups of 10 subjects. These steps are detailed next.

Initial Stimuli Selection (Original)

As already mentioned, the previous case study provided us with annotations to give us a basis to determine appropriate labels on. A total of 160 clips were rated. For this study, the rationale for selecting suitable stimuli considered three potential factors. First, speech clips can be selected based on the number of highest ratings received, and the more ratings a clip receives the more likely it has sufficient statistical power. For all speech clips, the number of ratings received ranged from 7 to 17 ratings, with 115 clips receiving 10 ratings or more. Second, the speech clips should, in some manner, be selected according to where the ratings were received on each emotion scale (Activation and Evaluation) since this would reflect on RQ2—to only select clips that received neutral ratings would likely be futile when making any comparisons. Third, one can also prioritise speech clips according to the level of agreement among raters, i.e. the label's statistical reliability. By considering all possibilities, the latter was concluded to be the best compromise. By arranging speech clips in order of best agreement levels, a sufficient number of ratings for each clip (ranging from 7 to 15) were retained, covering an even spread over each scale for each mean value obtained, and at the same time prioritising the label's reliability (see Appendix G for results on selected clips).

Agreement and reliability measures are achieved in several ways. Initially, the aim was to determine each clip's agreement level by calculating the standard deviation (*SD*) of the ratings received for each clip. However, because the number of ratings received for each clip differed, the measure of agreement according to the interval size at the clip's 95% confidence interval, or in fact, according to the size of margin error, was opted for. For instance, a margin of error of 0 would indicate complete agreement. Using the t-statistic (t) as the sample size was less than 30, the margin of error was calculated as:

$$E = t(\frac{\sigma}{\sqrt{n}}) \tag{8.1}$$

Speech clips were sorted in ascending order of agreement for both scales, each scale separately rather than combined. For the Evaluation scale, the interval size for all 160 speech clips ranged

from 0.335 to 3.49. The chosen speech clips for this experiment were within the range of 0.335 to 0.766. For the Activation scale, the overall range was from 0 to 3.35, and for the selected clips it ranged from 0 to 0.81. From each scale, the top 18 clips were selected according to their agreement level giving in total 36 clips (18x2). Overall, the clips had an interval range less than 0.81 on the Activation scale and less than 0.77 on the Evaluation scale. One must bear in mind that although ratings for each scale were obtained simultaneously, a speech clip that received high agreement on one scale does not necessarily receive similar values on the other. For example, one speech clip received an interval size of 0 (perfect agreement) on the Activation scale but a value of 0.989 on the Evaluation scale. Similarly, if a label is rated as emotionally salient on one scale, it may not be the same for the other scale. Out of 36 clips, 21 clips were considered that approximated as non-neutral on the Evaluation scale (greater or smaller than 2 plus or minus largest margin of error), and 29 as non-average on the activation scale. Incidentally, 16 of these clips were labelled as 'before' and 20 as 'after'[1].

Provision of filtered stimuli: 2 listening tasks

Two preliminary surveys, using two independent groups of 10 subjects, were carried out to (1) determine a suitable filtering condition to administer, and (2) to ensure that there was no comprehension of the spoken dialogue in the selected filtering condition (see Appendix H for survey and results). Both surveys were carried out in college labs to ensure that the task was carried out free from any external distractions. A laptop was used to present the audio in random order to the subject. The speech clips were listened to using Beyerdynamic DT 150 headphones, with a frequency response of 5Hz-30kHz. Due to the nature of the stimuli, it is essential to optimise the accuracy of the presented auditory stimuli. The characteristics of the DT 150—a closed headphone design and wide frequency response—attenuates unwanted ambient noise and ensures the sounds are delivered with adequate clarity.

Participants were informed to write down any words that they could comprehend, or give another free response feedback if so desired. For the first survey, subjects were presented with 18

[1] See Chapter 7 where clips were separated into two categories.

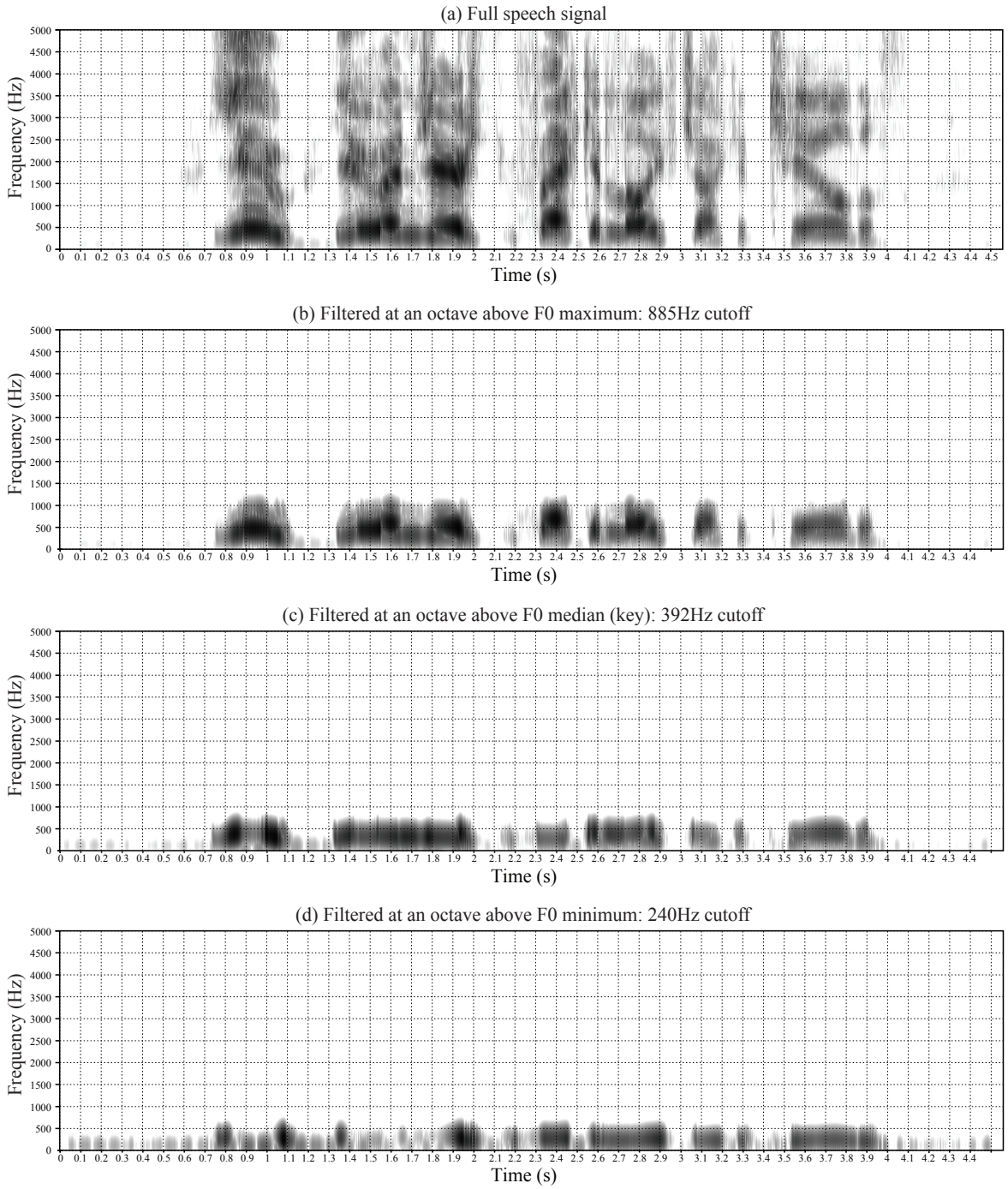


Figure 8.4: Spectrogram of example speech clip in its original form (a) and 3 filtering conditions (b), (c) and (d), where condition (c) is the final applied filtering measure.

different speech clips, 6 clips for each of 3 low-pass filtering conditions. The low-pass filtering cutoff points for the three conditions were set an octave above F_0 min, F_0 median (key)[2], and F_0 max. Figure 8.4 shows an example of a clip's spectrograms in its original condition (a) and its respective filtered conditions according to the aforementioned cutoff points (b), (c) and (d).

[2]The median is used here instead of the mean as it is more robust to deviating F_0 values.

The results were somewhat unexpected. Two speech clips were partly comprehensible in the filtered key condition but none in the F_0 max condition. As expected, participants could not comprehend any of the F_0 minimum conditions. Several participants remarked that the stimuli did not sound like human speech, but rather like a "rumbling" noise, music or rhythmic pulses. It was surmised, therefore, that filtering an octave above the F_0 minimum was excessively low and created inapplicable stimuli. Filtering an octave above the key is evidently less likely to be comprehensible compared to an octave above F_0 max because the cutoff point is at a lower value. Therefore, it was concluded that the key value of the clip as the filtering reference point was most suitable.

In the second survey, all 36 speech clips were low-pass filtered proportional to the speech clip's key. The 36 filtered speech clips were presented to 10 participants, and some dialogue was correctly perceived in 3 speech clips[3]. Moreover, 1 clip contained some low-frequency noise that became more perceptible when filtering was applied. In total, 4 speech clips were excluded from the main experiment, reducing the final number of speech clips to 32 for this experiment.

Final Stimuli Selection

Altogether, a total of 64 speech clips (32 x 2 for each condition) were used in the experiment. These clips were of short length ($\sim$5 seconds) assuming that there were no transitions in the emotional content. The amplitudes of the original speech clips were normalised at -0.3dB, while the filtered clips were normalised at -0.1dB.

As mentioned above (see section 4.2), most studies that use low-pass filtering to mask linguistic content use one or more *fixed* values as cutoff points. However, different speakers—and the extracted speech segments—vary in F_0 ranges and spectrum energy distribution. Therefore, a fixed filtering condition across all speech clips could potentially give different degrees of

[3]These clips received high values for Activation (3.76, 4, and 3.62), and were all labelled as 'after' in the previous case study

intelligibility. For this reason, the aim was to make the level of unintelligibility uniform across all speech clips. Accordingly, a *unique* filter cutoff value was exerted that was proportional to the parameters of the respective speech sample. As determined from the preliminary study, each clip was filtered with a cutoff value proportional to its F_0 median (see the example in Figure 8.4 (c)). The actual cutoff frequencies chosen were an *octave* above the clip's key[4] (F_0 median x 2), which ranged from 197Hz to 1162Hz for all 32 speech clips (see Appendix I for clip parameters). The pitch floor and ceiling settings were automatically adjusted relative to the clip's F_0 quantile values, which gives a better estimation of pitch extrema [448]. The 32 speech clips were low-pass filtered (Hann window), with smoothing at 20 Hz, using PRAAT 5.3.13 [449] software (see Appendix J for the script).

8.3.3 Online Rating Tool

As mentioned, the tool used for this experiment is based on the tool developed for the case study in the previous chapter. In Chapter 6, the changes that were made to the tool specific for this task were outlined. Minor changes were made to all pages based on the feedback and experience during the case study. Specifically, changes were made to the back-end design in order to facilitate the design of the experiment, as illustrated above (section 8.3.1), i.e. a within subject, 2-period crossover design. For example, if a participant was initially assigned to Group A, and was presented the non-filtered condition in the initial task, the participant would be presented with the filtered condition for the alternate task (two weeks later), and vice versa for Group B.

8.3.4 Selection of Subjects

For this experiment, 57 participants took part completing the two phases of the experiment, i.e. rating both conditions. Of these participants, mostly were newly recruited, with less than 10% that also took part in the previous study. Participants were asked if they had any hearing im-

[4]To obtain the key (F_0 median) value, a PRAAT script based on Celine De Loozes, 'Get_Speakers_register.praat', was used. This script minimises possible pitch tracking errors [448]. It can be found at: `http://www.celinedelooze.com/MyHomePage/Praat.html`.

pairments, and it was decided to exclude the ratings from those who did. Several demographic variables that could demonstrate potential differences in the emotional judgement task were collected and taken into account. As a reminder, this study differentiates linguistic and tonal aspects of speech. Various studies indicate that listeners identifying emotion in speech have an attentional bias towards either linguistic or prosodic content. These individual differences have been shown to be relevant to culture [321], age [438], gender [439], and handedness (or laterality) [440].

Native versus Non-native Speakers

As with the previous case study (Chapter 7.2.3), information on native and non-native English speakers was documented. Although the results in the case study revealed no significant differences, they may not apply in this study due to the nature of the task, which emphasises the prosodic aspects of speech. All participants were fluent English speakers, of which 8 did not have English as their native tongue—group sizes were of 49 and 8. All participants were presented with a release form prior to the experiment (see Appendix K), in which they were informed to only proceed with the task if they consented with the experiment requirements. The form and the experimental conditions of this experiment were reviewed and passed by the ethics committee at the Dublin Institute of Technology.

Age

Age was documented as age-related hearing loss needs to be taken into account, although age affects high frequencies more than low frequencies. Besides, there are ongoing studies on age-related differences in the perception and meaning of emotional stimuli. For example, the study by Grühn and Smith [438] revealed significant differences on a large number of German adjectives, especially on the valence—synonymous with Evaluation—dimension. Age ranged from 18 to 65 years, with one group consisting of 45 participants whose ages were 18 to 40, and the second group consisting of 12 participants who were 40 and above.

Sex

As regards sex, several studies have reported differences in emotional prosody processing between male and females [450, 439, 96]. The study by Schirmer and Kotz [451], for example, suggests that in contrast to females, men process linguistic and prosodic—or acoustic—information independently. Participants consisted of 30 males and 27 females.

Handedness

Finally, there is a considerable amount of research and debate on the subject of laterality. Various studies indicate that the two aspects of speech—linguistic and acoustic—are lateralised in the left and right cerebral hemispheres, respectively. The processing of nonverbal stimuli such as emotional prosody is relatively right-hemisphere lateralised, whereas the processing of articulate, verbal information is relatively left hemisphere lateralised [330, 440]. Similarly, studies show that there are individual differences in listeners perceiving complex sounds expressed in a left (F_0, or *synthetic* listeners) or right (spectral, or *analytical* listeners) hemisphere dominance. A study by Nilsenová et al. [440], for example, explored how listeners differed in identifying emotion according to synthetic and analytic listening modes. They classified listener preference using a pitch discrimination task and showed that *spectral* listeners performed better in an emotion judgement task. Because some listeners are more sensitive to overall spectral information, it is conceivable that certain listening modes may influence the perception of speech in this experiment for stimuli where higher frequencies are missing. Therefore, the rater's listening preference—operationalised as handedness—is also observed. Although notably uneven, 8 of the participants were left-handed and 49 were right handed.

8.4 Results

The statistical analysis reported in this section were conducted using the SPSS statistical package [452]. The duration of the rating accumulation process over the two phases was 9 weeks,

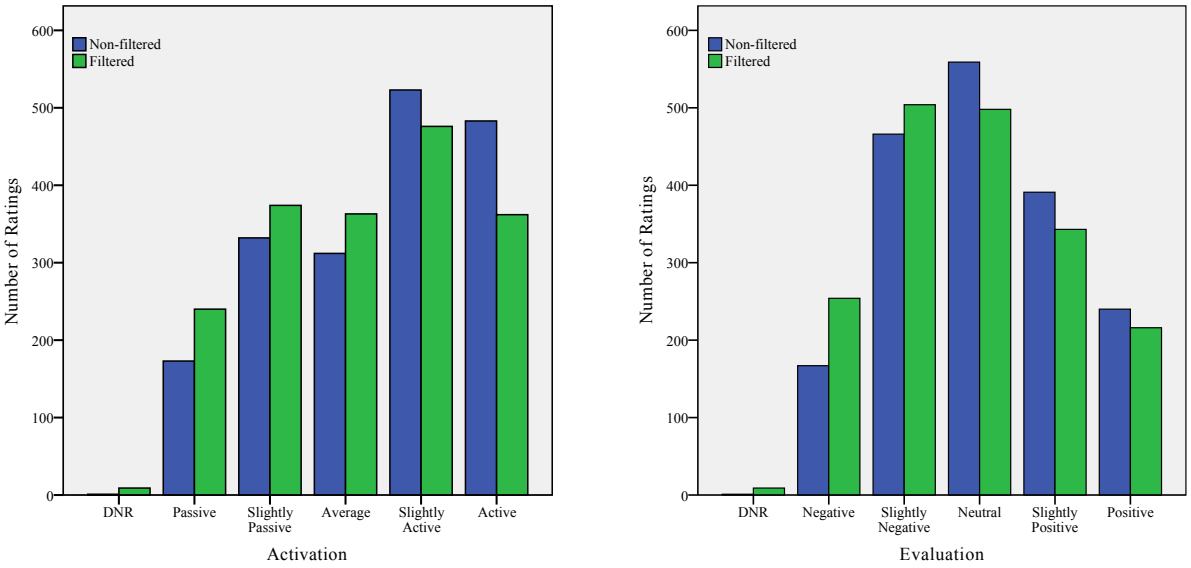


Figure 8.5: Distribution of the ratings received for each condition, non-filtered and filtered speech—for the Activation (left) and Evaluation (right) scales. DNR = "Do Not Rate".

in which all 57 participants who took part had listened and rated 64 speech clips (32 for each condition). A total of 1823 ratings were received for each scale in the *non-filtered* condition. Only 1 DNR (Do not rate) was received[5], which amounts to only 0.05% of the maximum number (57x32 clips) of potential ratings. For the *filtered* condition 1815 ratings for each scale, and 9 DNR ratings, were received. As expected, the 9 DNR showed the filtered condition was harder to rate; however, it only amounts to 0.49% of the maximum number of potential ratings. In both conditions, the percentage of DNRs was small, which showed that there was little uncertainty in the participants rating either condition. The DNR ratings were deemed as insignificant and, therefore, disregarded any further analysis on the DNR ratings—these values were incorporated as 'missing values'.

For each scale, Figure 8.5 shows the number of ratings received for each class in both conditions. The mean value on the *Activation* scale for the *non-filtered* condition was 2.44 (*SD*=1.308) with slight negative skewness. The data on this scale is spread over all classes, but the values appear most often in the Slightly Active and Active classes, respectively—this is similar to the trend in the case study (see Chapter 7, Figure 7.3). The mean value on the *Activation* scale for

[5]If a participant chose "Do not rate", the value DNR was written to the database for both scales.

the *filtered* condition was 2.19 (*SD*=1.327) with Slight Negative skewness. The data on this scale is spread more evenly compared to the non-filtered condition, but again, the values occur most often in the Slightly Active class.

For the Evaluation scale, the non-filtered condition had a mean value of 2.04 (*SD*=1.167) and a slight positive skew. Although ratings were spread over all classes, the data occur most frequently in the neutral class, as somewhat expected. The ratings for this scale reveal a similar trend to the case study. The Evaluation scale in the filtered condition had a mean value of 1.87 (*SD*=1.219), again with a slight positive skew. Although most data for this condition prevail in the Slightly Active class, the Neutral class peaks on par with it.

8.4.1 Participant Demographics

As already mentioned, demographic information from participants was gathered that was assumed to be potentially significant, especially for the prosodic aspects of the study—filtered condition. This included English as a first language (nativeness), age, gender, and handedness. To investigate if there is an interaction between the demographic variables (between-subjects factor) and the stimuli conditions (within-subjects factor) on the ratings on each scale (dependent variable), the most suitable statistical analysis is the mixed-design ANOVA. On the obtained data, preliminary analysis indicated violations of the assumption of normality, as assessed by Shapiro-Wilk's test ($p > 0.05$). Although, the mixed ANOVA is somewhat robust to deviations from normality, some of the analysis showed there was no homogeneity of variances, as assessed by Levene's test of Homogeneity of Variance ($p > 0.05$). In spite of the violations of the assumption of normality, the results for the mixed ANOVA are presented in Appendix L.

In this section, non-parametric analysis are reported on. Unfortunately, there is no well matched non-parametric alternative to the mixed ANOVA, differences between groups on the overall obtained ratings (combined conditions), and independently for each condition, non-filtered and

		ACTIVATION						EVALUATION					
		z-value	*p*-value	*r*	Md_A	Md_B	*U*	*z*-value	*p*-value	*r*	Md_A	Md_B	*U*
Nativeness	Non-filtered	-0.075	0.941	0.002	2.5	3	108063	-1.596	0.11	0.037	2	2	99586
(A=Non-native	Filtered	-1.596	0.11	0.037	3	2	99586	-0.681	0.496	0.016	2	2	103415
B=Native)	Overall	-0.858	0.391	0.014	3	3	417791	-0.639	0.523	0.011	2	2	421279
Handedness	Non-filtered	-0.196	0.845	0.005	3	3	177683	-0.839	0.401	0.020	2	2	173080
(A=Right	Filtered	-1.108	0.268	0.026	2	2	170238	-1.793	0.073	0.042	2	2	165379
B=Left)	Overall	-0.658	0.511	0.011	3	2	701207	-1.836	0.066	0.030	2	2	677401
Gender	Non-filtered	-1.244	0.214	0.029	3	3	385215	-0.153	0.879	0.004	2	2	396783
(A=Male	Filtered	-1.744	0.081	0.041	2	2	376553	-0.661	0.509	0.015	2	2	388049
B=Female)	Overall	-1.238	0.216	0.021	3	3	1611774	-0.82	0.12	0.014	2	2	1624746
Age	Non-filtered	-3.2	0.749	0.075	3	3	289720	-1.612	0.107	0.038	2	2	277900
(A=18-40	Filtered	-2.439	***0.015***	0.057	2	3	266729	-1.62	0.105	0.038	2	2	274273
B=>40)	Overall	-1.531	0.126	0.025	3	3	1123676	-2.269	***0.023***	0.038	2	2	1104754

Table 8.1: Mann-Whitney U test results; Md = median.

filtered, are investigated. For this, a multiple Mann-Whitney *U* tests is performed and the results are listed in Table 8.1. The only demographic variable that indicated a significant difference was that of age. A significant difference is observed on the Activation scale under the filtered condition ($p = 0.015$, $r = 0.057$), and on the overall obtained ratings on the Evaluation scale ($p = 0.023$, $r = 0.038$). Both differences indicated a very small effect size.

8.4.2 Inter-rater Measures

As in the previous chapter (see section 7.3.4), two inter-rater measurements are performed: Krippendorff's α (inter-rater reliability) and standard deviation measurements (inter-rater agreement). Out of the 64 speech clips (32 for each condition) that were rated, 1 DNR rating was received in the non-filtered condition, and 9 DNR ratings for the filtered condition. In the following, speech clips that were rated as DNR were treated as 'missing' values.

Inter-rater Reliability

Table 8.2 shows the Krippendorff's α [17] (ordinal scale) achieved on each scale for each stimuli condition. As shown, the agreement coefficients are higher for the Activation scale, with the highest α score observed on the Activation scale, non-filtered condition (0.588), and the lowest on the Evaluation scale also for the non-filtered condition (0.26). The observed α for the Evaluation scale is slightly higher for the filtered condition than the non-filtered condition.

	Non-filtered		Filtered	
	α	Mean *SD*	α	Mean *SD*
Activation	0.588	0.8356	0.555	0.8602
Evaluation	0.26	0.9399	0.294	0.9841

Table 8.2: Krippendorff's α [17], and mean standard deviation (SD) values for both conditions on both scales.

Inter-rater Agreement on Individual Speech Clips

As mentioned in the previous chapter, to assess agreement levels on *individual* speech clips, one can use the standard deviation (SD) of the ratings received for the respective clip[6]. Table 8.2 shows the average standard deviation values received for both scales in both conditions. Further details on correlation and mean difference measurements of the obtained SD values are included in the next section.

In the previous Chapter, the median values of all speech clips where compared against the obtained SD measures using a stacked bar chart (see Figure 7.6, section 7.3.4). It allowed us to visually compare SD ranges against the tendency of the obtained median values. In other words, it allowed us to determine agreement measures that are exhibited for a particular class. For example, it may be expected that high agreement is only affiliated with non-emotional speech, i.e. Neutral (on the Evaluation scale) or Average (on the Activation scale) speech. SD measures ranged from 0.13 to 1.48 for both conditions on both scales, which was subdivided into 4 ranges.

Figure 8.6 shows the distribution according to the SD range (stacks) and the obtained median values (scale classes). For the Activation scale (top left) in the non-filtered condition, the diagram demonstrates that the highest agreements (SD range of 0.0 to 0.5) obtained were mostly for speech clips with a median value on the Active class, and that it also obtained a large num-

[6]In this case, the Confidence Interval range was not used because ratings were evenly proportioned across all speech clips and, therefore, did not need to consider the mostly varying numbers of received ratings—as done for the ratings obtained from the case study.

ber for those with an SD range of 0.5 to 1.0, similar in number to those received for the Average class. For the filtered condition (top right), speech clips whose median values were in the Active class again obtained the highest agreement, but speech clips with an SD range of 0.5-1.0 were mostly in the Slightly Passive class. On the Evaluation scale, neither condition obtained an SD value in the range of 0.0-0.5. For the non-filtered condition (bottom left), the highest obtained scores manifested mainly for those clips whose median values were Neutral. For the filtered condition (bottom right), however, the highest SD range occurred to the same degree in both Slightly Negative and Neutral classes.

8.4.3 Associations and Group Differences

In addition to agreement measures, this section examines both the correlations and mean differences between ratings from the two different conditions. Correlation measures are concerned with the *relative* similarity of the ordering of ratings between both conditions, but they are not affected by and do not demonstrate any *absolute* differences. To demonstrate the significance of reporting both, let us consider an example of ratings provided in Figure 8.7. The ratings are illustrated for the original condition (a) and for 3 potential ratings (b), (c), and (d) from the filter condition. Visually, it suggests some potentially interesting comparisons. Ratings from (c) and (d), i.e. the filtered condition, both demonstrate a significant difference with that of (a)—Table 8.3 shows a rating difference of 0.89 between (a) and (c), and a rating difference of 2.11 between (a) and (d). The diagram shows that (a) and (c) exhibit a very similar trend. One may suggest, for example, that the scale seems to be restricted for the filtered condition, as clips 4, 5, and 6 all receive a rating of 0. By visually comparing (a) and (c) it suggests that filtered speech is rated *relatively* similar but with a decrease on its scale. This similarity is mirrored in a Pearson correlation coefficient of 0.976 (see Table 8.3), while the between-group test, such as the t-test and the Wilcoxon Signed Rank Test, indicates a statistically significant decrease in its rating. The results in (b) demonstrate no significant mean difference, but do demonstrate a perfect negative correlation that should not be overlooked. Similarly, the significant decrease of the mean difference—of a considerable size—between (a) and (d) should not be ignored even though it has a low correlation measure. With this in mind, correlation measurements and

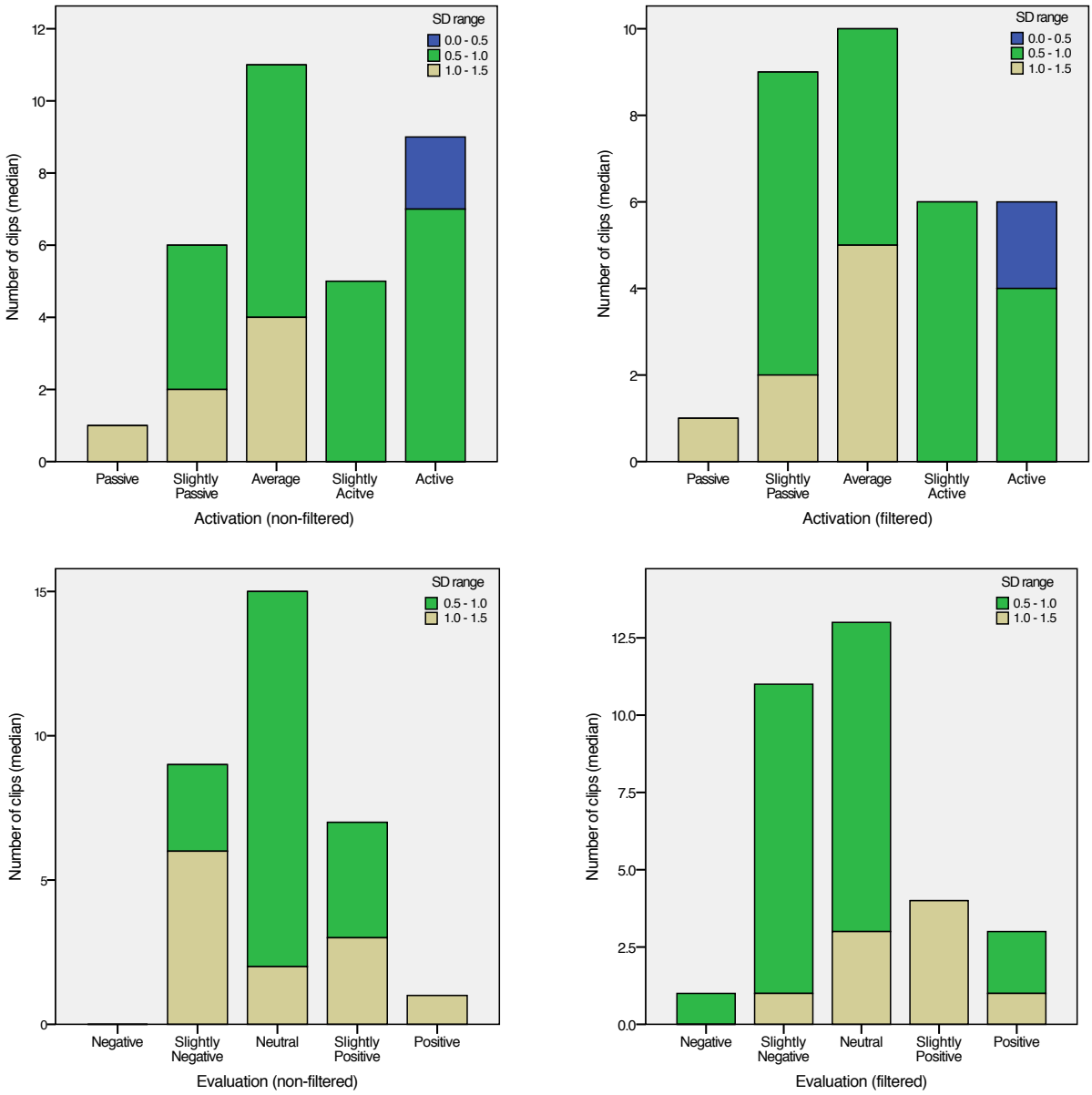


Figure 8.6: Distribution of speech clips with respect to the SD value and the median value obtained.

paired differences between are calculated for (1) ratings at the individual level and at the group level (mean), and (2) the *SD* measurements—interpreted as agreement—on each individual clip.

Correlation Measures

For non-parametric measures, both Spearman's rank correlation (ρ) and Kendall rank correlation coefficient (τ) are used to measure a pairwise correlation between two ordinal level

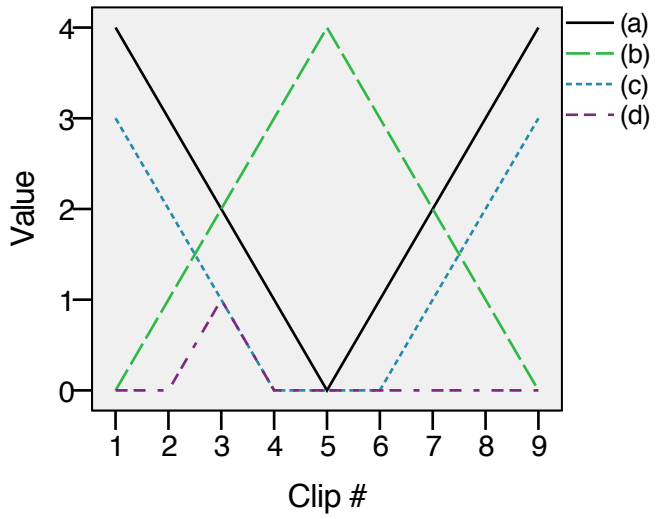


Figure 8.7: Example of different ratings; (a) data from the original speech clip; (b), (c), and (d) are other possible outcomes for the filtered conditions.

	Correlation	Mean difference
(b)	-1.00	NS
(c)	0.976	S: -0.89
(d)	0.06	S: -2.11

Table 8.3: Association and group differences tests for condition (a) against (b), (c), and (d).

variables—both of which capture a monotonic relationship[7]. Both Spearman's ρ and Kendall's tau (τ) measures range between -1 and $+1$. No association is signified with a value of 0, a perfect negative association is indicated with a value of -1, and a perfect positive association is indicated with a value of $+1$. However, Kendall's τ and Spearman's ρ imply different interpretations in the correlation value[8]. A significant aspect of the Spearman's ρ is that it involves squaring the deviations. In comparison, Kendall's τ is less sensitive to large discrepancies among a small number of ratings. In other words, it would be less sensitive when only one instance has a large deviation between the two conditions, but the remainders are otherwise perfectly concordant—it would, therefore, be less sensitive to one or two dubious ratings.

Spearman's ρ first computational procedure assigns ranking scores to the values of two different variables. Following that, the procedure involves calculating the differences between two successive rank values for a number of individual items (n), and then squaring the resulting deviations (d^2). The formula for Spearman's ρ is as follows:

[7]A monotonic relationship is a causal relationship that shows that either both variable values increase together, or as one increases the other decreases. The relationship preserves the given order but is not necessarily linear.

[8]In most cases, Spearman's ρ will be larger than Kendall's τ, but this is not always the case. If the deviations are bigger in a smaller number of cases, a smaller Spearman's ρ will be obtained.

$$\rho = 1 - \frac{6(\sum d^2)}{n(n^2-1)} \tag{8.2}$$

Kendall's tau τ represents the degree of concordance between two variables. It is a measurement of the number of concordant pairs (C) versus the number of discordant pairs (D). There are three variants for Kendall's tau, τ_a, τ_b, and τ_c, from which Kendall's τ_b is reported on here because this statistic, unlike Kendall's τ_a, makes adjustments for tied ranks—τ_b and τ_c only differ in the way they handle rank ties. The formula for Kendall's τ_b is:

$$\tau_b = \frac{C-D}{\sqrt{(C+D+Y_0)(C+D+X_0)}} \tag{8.3}$$

where X_0 is the number of pairs not tied on X, and Y_0 is the number of pairs not tied on Y [453].

Correlation between Individual Ratings

First, the ordinal correlation is evaluated of the ratings received for non-filtered speech and its corresponding filtered speech in terms of Kendall's τ_b. Kendall's τ_b is calculated between each rating for the non-filtered and filtered speech clips. The results for Activation (τ_b = 0.469, N = 1814, 2-tailed, p <0.0005) indicate there was a moderate positive correlation between the Activation received for the non-filtered speech and the filtered speech.

In addition, the correlation for each participant on the Activation scale is calculated. Of the 57 participants, the correlation was small (0.1 $<\tau_b \leq$ 0.3) for 7 participants, while for 19 participants the correlation was moderate (0.3 $<\tau_b \leq$ 0.5), and strong (0.5 $<r \leq$ 1.00) for the remaining 31 participants. For the analysis of Evaluation, the results for each rating (τ_b = 0.144, N = 1814, 2-tailed, p <0.0005) show there was a small positive correlation between the Evaluation perceived in the non-filtered speech and the filtered speech. The correlation obtained for each participant on the Evaluation scale showed that there was a small negative correlation (0 $<\tau_b$ $\leq$ -0.3) for 14 participants, a small positive correlation for 31 participants (-0.3 $<\tau_b \leq$ 0.0), and a moderate positive correlation (0.3 $<\tau_b \leq$ 0.5) for the remaining 12 participants.

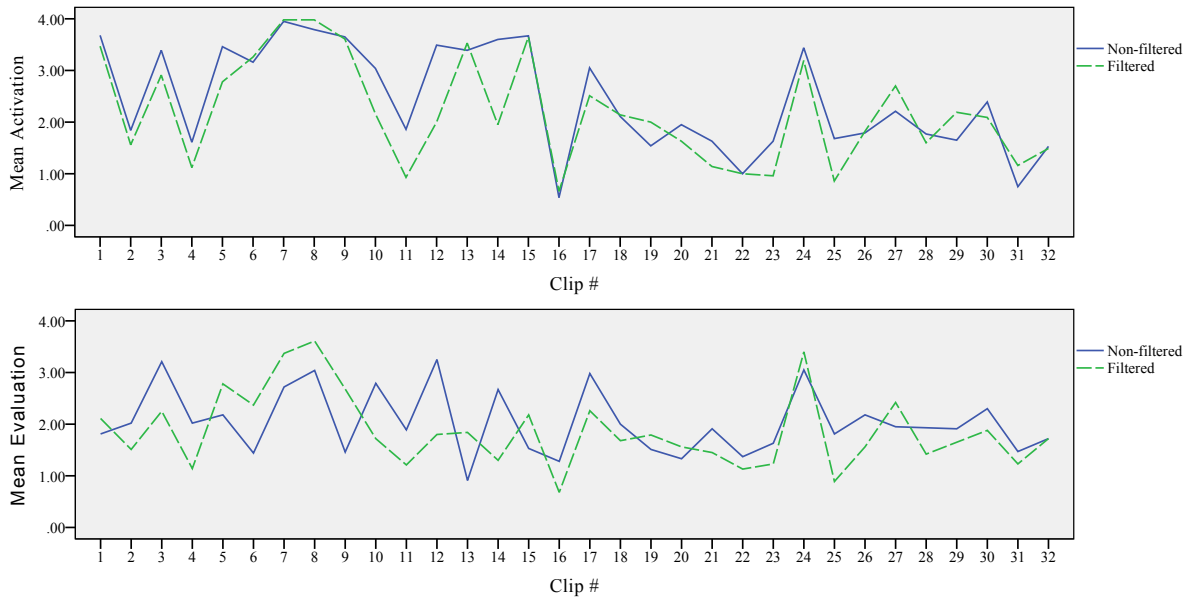


Figure 8.8: Mean values obtained for each clip for Activation (above) and Evaluation (below). For the Evaluation scale, 0=Negative, 1= Slightly Negative, 2=Neutral, 3=Slightly Positive, 4= Positive. For the Activation scale: 0= Passive, 1=Slightly Passive, 2=Average, 3=Slightly Active, 4=Active.

The correlation values between each rating for each scale were compared with the Spearman's ρ correlation coefficient. In this case, Spearman's ρ yielded similar—but slightly higher—values for the Activation scale ($\rho = 0.555$, $N = 1814$, 2-tailed, $p < 0.0005$) and on the Evaluation scale ($\rho = 0.170$, $N = 1814$, 2-tailed, $p < 0.0005$).

Correlation between Mean Values

As well as measuring the correlation between individual ratings, Kendall's τ_b for the mean values for each clip is calculated between the non-filtered and filtered conditions (see Figure M.1, Appendix M for Mean scatter plots). Figure 8.8 shows the obtained mean values for each clip on each scale, for both conditions. The mean values for each clip on the Activation scale ($\tau_b = 0.660$, $N = 32$, 2-tailed, $p < 0.0005$) show that there was a strong positive correlation between the mean values for Activation received for the non-filtered speech and the filtered speech. For the mean values on the Evaluation scale, the results for each rating ($\tau_b = 0.270$, $N = 32$, 2-tailed, $p = 0.031$) show there was a small to moderate positive correlation between the

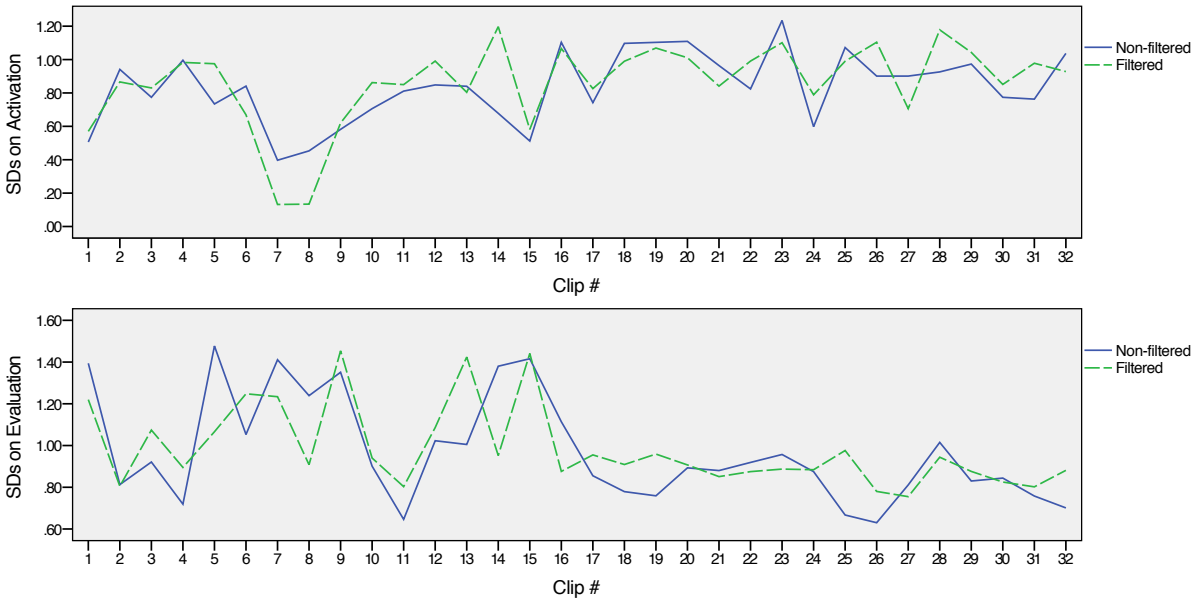


Figure 8.9: Standard deviation (SD) values for individual speech clips for Activation (above) and Evaluation (below) scales 0.

mean values for each clip on the Evaluation scale perceived in the non-filtered speech and the filtered speech. This shows that the correlation increased when the ratings were averaged for each speech clip.

Once more, the results with the Spearman's ρ correlation coefficient were confirmed. For the mean values on the Activation scale, the results (ρ = 0.836, N = 32, 2-tailed, p <0.0005) indicated a strong positive correlation. For the mean values on the Evaluation scale, the results (ρ = 0.384, N = 32, 2-tailed, p <0.0005) indicated a moderate positive correlation (see Figure M.2, Appendix M for correlation diagrams).

Correlation between Standard Deviations

In addition to observing the agreement measures (*SD*) on individual speech clips (section 8.4.2), the correlation values were calculated between the agreement measures (*SD*) obtained for each clip in both conditions, using Pearson product-moment correlation coefficient (r). Figure 8.9 illustrates the received SD values for each clip for the non-filtered and filtered condition, showing somewhat inconsistent SD values. The obtained *SD* values, however, appear to follow a

similar trend between the two conditions (see Figure M.2, Appendix M for SD scatterplot). For the Activation scale, there was a very strong positive correlation for the SD values between both conditions, $r = 0.723$, $n = 32$, $p < 0.0005$; and there was a strong positive relationship, $r = 0.663$, $n = 32$, $p < 0.0005$, between the SD values for the Evaluation scale.

Differences between Individual Ratings

As mentioned above, correlation measures do not provide us with any information on *absolute* differences in the height of the ratings given in different conditions. In fact, a perfect correlation will exist if the ordering of ratings between both conditions is exactly the same, but this does not inform us of any differences.

To investigate the differences in the perception of Activation and Evaluation on the individual level between the two conditions (non-filtered and filtered), the ratings were subjected to the nonparametric Wilcoxon Signed Rank Test. The Wilcoxon Signed Rank Test revealed a statistically significant decrease in the level of Activation rated in the filtered condition compared to the non-filtered condition, $z = -8.42$, $p < 0.001$, with a small effect size ($r = 0.14$). The median Activation rating for the filtered speech clips ($Md = 2$ = Neutral) was lower than the Activation rating for the non-filtered clips ($Md = 3$ = Slightly Active). We can observe (Table 8.4) that there are more speech clips with a median value for the *non-filtered* Active class than there are for the *filtered* Active class, but more instances in the *filtered* Passive class than the *non-filtered* Passive class. Similarly, we can observe that more instances of the mode value appear in the Slightly Passive class, and Slightly Active class.

For the Evaluation scale, statistically significant decrease was found in the level of Evaluation perceived in the filtered condition, $z = -4.833$, $p < 0.001$. The effect size was small ($r = 0.08$). The median of Evaluation for the non-filtered speech clips ($Md = 2$ = Neutral) was the same for the overall median of Evaluation for the filtered clips ($Md = 2$ = Neutral). Table 8.4 shows that there are more instances of speech clips with median values in the extreme classes (1

	Non-filtered		Filtered	
Activation	Md	M	Md	M
Passive	1	4	1	5
Slightly Passive	6	7	9	7
Average	11	5	10	5
Slightly Active	5	6	6	9
Active	9	10	6	6

	Non-filtered		Filtered	
Evaluation	Md	M	Md	M
Negative	0	0	1	2
Slightly Negative	9	11	11	11
Neutral	15	13	13	9
Slightly Positive	7	3	4	6
Positive	1	5	3	4

Table 8.4: The number of speech clips in each class with respective median (Md) and mode (M) values for the non-filtered and filtered conditions—for the Activation (left) and Evaluation (right) scales.

Negative and 3 Positive) for the filtered condition. For the mode values in the filtered condition, the ratings occur more frequently in the Negative and Slightly Positive class, but less for the Neutral and Positive class.

Differences between Mean Values

Measuring differences at the individual level may give different results to measuring differences at the group level, i.e. in terms of a clip's obtained mean value. Because mean values are continuous, a paired-sample t-test is conducted to determine whether the mean difference between the mean values for each clip observed for both conditions is significantly different. A paired-sample t-test was conducted for both scales, and it showed, for the Activation scale, a statistically significant difference between the means of each clip in the non-filtered condition (M = 2.45, SD = 1.01) and the filtered condition (M = 2.19, SD = 1.00), $t(31)$ = -0.2.777, p <0.009 (two-tailed). The mean decrease was 0.26 with a 95% confidence interval ranging from 0.068 to 0.44. The eta squared statistic (0.2) indicated a small effect size. For the Evaluation scale, it showed no statistical significant difference between the means of each clip in the non-filtered condition (M = 2.04, SD = 0.625) and the filtered condition (M = 1.87, SD = 0.72), $t(31)$ = 1.379, p <0.178 (two-tailed).

Differences between Standard Deviation Values

Figure 8.9 illustrates a similar trend in SD values obtained for speech clips in both conditions. In fact, section 8.4.3 showed that there was a very strong correlation on the Activation scale

and a strong correlation on the Evaluation scale. To determine whether the mean differences of the SD values are statistically significant, a paired-sample t-test is performed on the obtained SD values to evaluate the agreement levels of each clip for each scale. The paired sample t-test showed no statistically significant difference for the Activation scale (M = 0.84, SD = 0.21) in the non-filtered condition and the filtered condition (M = 0.86, SD = 0.24), $t(31)$ = -0.806, p <0.426 (two-tailed). Similarly, there was no statistically significant difference for the Evaluation scale in the non-filtered condition (M = 0.97, SD = 0.25) and the filtered condition (M = 0.98, SD = 0.19) $t(31)$ = -0.424, p <0.675 (two-tailed).

8.5 Discussion

Demographic information did not show that there were any considerable differences between the selected groups and how they rated non-filtered and filtered speech. Demographic information did not appear to influence the performance of raters so all participants were included in the analysis for inter-rater measures, associations and mean differences.

Distribution of Ratings

The distribution of the ratings on both scales for the two conditions is shown in Figure 8.5. A slight negative skew is observed for the Activation scale in both conditions, and a slight positive skew for the Evaluation scale in both conditions. The non-filtered condition contains more ratings in the Active classes for the Activation scale, and more instances in the Positive classes for the Evaluation scale, which may be somewhat consistent with the assumption that upper frequencies lead to a loss of certain emotional cues [146]. However, for the filtered condition, there is only a slight increase in the number of ratings received for the Average class and, in fact, a slight decrease in the number of ratings received for the Neutral class. For the Activation scale, there is an increase in the number of Passive ratings. Similarly, there is an increase of Negative ratings for the Evaluation scale. This may suggest that the loss of upper frequencies, or linguistic content, is rated as more Negative (see also [145]) and Passive.

Inter-rater Measures

Inter-rater reliability for the emotional judgement task was rather low for Activation in both conditions where α was 0.588 for the non-filtered condition and 0.555 for the filtered condition. More so, the Evaluation scale received an α of 0.26 for the non-filtered condition and an α of 0.294 for the filtered condition. While some works report on a decrease in inter-rater reliability on filtered speech [146, 326], these results suggest this is only the case for the Activation scale. It was acknowledged in the previous chapter (section 7.4.3) that low inter-rater reliability (α) and agreement measures (SD) are expected for non-filtered speech. Compared to other studies, the results suggest that reliably identifying emotion is a difficult task. One may expect that rating emotion from filtered speech may be somewhat more difficult. However, there are marginal differences in the reliability and agreement measurements between the two conditions. The highest observed α score (0.588) on the Activation scale for the non-filtered condition would be somewhat expected. The low α scores for Evaluation in both conditions show the difficulty of the task, whether or not the speech was filtered or not. It is clear that natural speech is inundated with ambiguity. In addition, the evaluator assesses emotional content in a speech clip differently to how the emotion is truly felt by the speaker. Factors such as *display rules*, *deception*, and *systematic ambiguity* (see [156] for an overview) play an important role in the dissent of how emotion is perceived. Linguistic cues may not always concur with the paralinguistic cues, and, when evaluating speech, participants may prioritise acoustics over semantics, or vice versa. For this reason, it may be expected that the filtered condition would receive a higher inter-rater score than the non-filtered condition due to the minimisation of interference from the linguistic cues. This is, however, difficult to determine because the α and SD agreement measures are giving contradictory interpretations. The inter-rater reliability value, as determined by α, is showing a small increase in reliability for the filtered condition, yet on the individual clip level, SD is showing a slight decrease in agreement (increase in SD).

Similarly, the standard deviation was measured in both conditions as a measure of agreement on an individual clip. The standard deviation values appear to follow a similar trend between

the two conditions (see Fig 8.9). In fact, the correlation measurement between the two conditions was very strong for the Activation scale and strong for the Evaluation scale. The strong correlation could suggest that the ambiguity and difficulty in rating certain speech clips is attributable to the acoustics alone, irrespective on semantic content. Although the SD values of clips 15 to 32 predominantly increase in value on the Activation scale, they appear to decrease in value on the Evaluation scale.

Association and Group Differences

The association of ratings between the two conditions was examined by calculating Kendall's τ_b over all individual cases— each individual rating of the non-filtered clip compared with the rating of the filtered clip. It showed that there was a moderate positive correlation for the Activation scale, and a small positive correlation for the Evaluation scale. The analysis on each participant showed that the majority (31) of the participants had a strong correlation between each condition for the Activation scale and a small positive correlation for the Evaluation scale. Interestingly, 14 participants showed a small *negative* correlation between the two conditions for the Evaluation scale. Again, this may be because the interpretation of paralinguistic cues may not correspond to the interpretation of linguistic cues. That is, emotions can be transmitted deceitfully, and speech that may be semantically negative may be expressed in a positive manner with, for example, laughter. In this case, linguistic and paralinguistic cues may have opposing impressions.

In addition to obtaining Kendall's τ_b on individual cases, Kendall's τ_b was calculated to compare the mean values for each clip in each condition. The graph (see Fig 8.8) appears to suggest that the ratings for each clip follow a broadly similar trend, especially for the Activation scale. The results indicated a strong positive correlation for the Activation scale and a small to moderate positive correlation for the Evaluation scale. This showed that the strength of correlation between the two conditions increased when comparing the mean values for each clip, i.e. the correlations were stronger at the group level for each clip as opposed to the individual level of

each rating—similar to the findings of Teshigawara et al. [289]. The strong correlation between the filtered and non-filtered conditions would be expected and in agreement with earlier findings [32]) isolating pitch/prosody-related features. The significant decrease in the perception level of Activation, similarly expected, may be explained by the removal of the high frequencies. Clips 12 and 14 have the biggest difference (1.65 and 1.49 classes, respectively) between their mean Activation values. Both speech clips contained laughter that may be a factor in the decrease in perception of Activation level in the filtered condition.

Although there is a small decrease in the level of Evaluation for the filtered condition, the overall results do not necessarily suggest a preference or importance in the emotion perception of lexical content, as there are fewer Neutral ratings in the filtered condition. The low correlation, however, does suggest incongruence between lexical and acoustical cues (cf. [320]). For the evaluation scale, clip 9 is rated more positive in the filtered condition compared to its counterpart in the non-filtered condition with a mean value of 1.46 in the non-filtered condition and 2.28 in the filtered condition. The spoken part in this clip "oh my God... we were doing so well", may be semantically perceived as Slightly Negative, although it is rated 'Slightly Positive' in the filtered condition. This clip may be an example where the speech clip's *acoustic* significance opposes the effect of its *semantic* meaning.

8.6 Conclusions

This chapter investigated the effect that low-pass filtering has on the perception of emotion—as described by Activation and Evaluation dimensions—in naturalistic speech, with the general aim of exploring the inference of emotion from nonverbal aspects of speech. The spread of the distribution (section 8.4) indicates that low-pass filtered speech (removing semantic content) does signal emotion in the Activation-Evaluation space, thus addressing research question four (RQ4). In relation to research question five (RQ5), it appears that the perception of Activation and Evaluation is influenced by low-pass filtering, but that the effect is relatively small. To some extent, it was expected that removing linguistic content would eliminate any incongru-

ence between the acoustic and linguistic channels, making it less ambiguous—which would be suggested by the inter-rater measures (section 8.5). However, although there was an increase in inter-rater reliability for the Evaluation scale in the filtered condition, this increase was relatively small. Moreover, there was a strong correlation between agreement measures, based on standard deviation, between both conditions, suggesting that the difficulty in rating speech clips can be attributable to the acoustics, and not the incongruence of both channels. In future research, it would be of interest to obtain a speech dataset that provides for systematically ambiguous, and opposing acoustic and semantic meaning. This, however, may not be straightforward with naturalistic speech. With acted speech, however, one could systematically generate speech with opposing paralinguistic and linguistic content, such as speech with negative semantic content expressed with positive affect e.g. sarcasm or irony. It may be conceived that higher inter-rater agreement could be achieved for filtered speech of an ambiguous nature when its linguistic cues are removed.

Overall, the work in this chapter demonstrated that low-pass filtering is a useful tool to remove semantic content while preserving salient prosodic cues. To this end, this chapter contributed to answering research questions RQ4 & RQ5.

9

Conclusion

9.1 Summary of work

This work is motivated by the long-term goal of developing systems to automatically recognise emotion from naturalistic speech. The area of focus for this thesis was the nature of perception of emotion from vocal acoustics. This research began by reviewing four main perspectives that attempt to conceptualise emotion (Chapter 2), giving a general overview of the most influential theories that these perspectives have arisen from. In this review chapter it was determined that emotion is commonly conceptualised as a multifaceted phenomenon that manifests itself in different components (section 2.5), with research focusing on individual or combined aspects of these components.

To investigate emotion scientifically, the next chapter of the review (Chapter 3) gave details on how to measure and assess emotion, and how to represent emotion in order to make emotional states distinguishable. Three independent assessment types were explained and discussed (section 3.2). While it is theorised that these measurable components cohere, the review demonstrates that there are contradictory findings also. Moreover, an important distinction between two emotion label descriptions was discussed, the labels being representative of speaker or listener realisation (section 3.2.5). Focusing on effect-type labels, the two approaches to selecting judges (section 3.2.6) were considered and it was decided to select judges whose expertise are not defined, rather than focusing solely on expert judges to rate the speech material. Following emotion assessment methods, the review documented how to represent and classify emotions (section 3.3). The literature suggests that representational frameworks can be broadly distinguished by discrete (section 3.3.1) or dimensional theory (section 3.3.2). The discrete representation approach indicates that there is no definitive (or optimal) list of emotions suitable for naturalistic speech. The discrete representations that were considered in the review were prototypical categories (section 3.3.1), such as the well-known "Big Six" (section 3.3.1), subordinate categories (section 3.3.1), and cover classes (section 3.3.1). The alternative approach that was considered was the dimensional approach (section 3.3.2), which represents either abstract underlying factors between emotional states (e.g Activation, Evaluation, and Control), or is based on appraisal of objects, events or situations (e.g. the OCC model). The literature on labelling methods for naturalistic data suggests that they are more complex than the methods previously used for acted material. In this regard, several in-depth descriptive schemes were reviewed, and what appears to be common in many descriptive schemes is the incorporation of the two-dimensional model: Activation and Evaluation (section 3.3.4). From this chapter of the review, it is suggested that the Activation and Evaluation dimensions are most suitable for the spontaneous emotional speech material provided for this investigation.

Following this, various aspects of the conceptual basis of expression and perception of emotion in vocal communication (section 4.1) were presented, which gave us a better understanding

of what is involved in labelling emotional speech. The Brunswikian Lens model of emotion (section 4.1.1) was introduced, and the different areas of study were delineated with regard to the various aspects of the communication process illustrated by the model (section 4.1.2-4.1.6). It was argued that such a conceptual framework is methodologically suitable for the study of vocal expression. Labelling methods based on perception tests were then considered, indicating that much work correlate acoustic features with labels that are derived from prosodic and semantic content (section 4.2). The different speech material types were presented, indicating differentiation based on the elicitation type (section 4.3). Emotion authenticity (section 4.3.4) and audio quality (section 4.3.5) issues were discussed and it was argued that mood induced speech provide for an appropriate source of material.

The final chapter of the review (Chapter 5) examined the most prevalent acoustic features studied in research of emotional speech, which include prosody, pitch, time, intensity, voice quality, and spectral-related features. The description of vocal acoustics as a two-stage process was first presented, involving the glottal energy source and vocal tract filtering (section 5.1). Four different conceptualisations of prosody were then considered, demonstrating that it is at the suprasegmental level, eminent in the communication of emotion (section 5.2). In this section, several software frameworks for transcribing prosodic features were presented. At the acoustic realisation of the prosodic phenomenon, several features were presented in more detail, including fundamental frequency (F_0) related measurements (section 5.3), time-related measurements (section 5.4), and intensity-related measurements (section 5.5). The literature indicates that these features are an important component for conveying non-lexical information such as emotion. Furthermore, the literature on voice quality suggests that it too is fundamental in the expression of emotion and is utilised by listeners to distinguish states on both Activation (Arousal) and Evaluation (Valence) dimensions (section 5.6). The final category of features considered was spectral features (section 5.7), which are measured for speech emotion recognition, including harmonics, formants, spectral tilt, mel-frequency cepstral coefficients, and long-term average spectrum analysis.

The provision of an appropriate emotional speech corpus was identified as the initial goal. The construction of an emotional speech corpus is a two-stage process, collecting suitable emotional speech data, and providing labels that describe its emotional content. By carrying out perception tests, the work carried out in this thesis addresses the latter. Because studies on naturalistic emotional speech are relatively limited, the development of this corpus is a major contribution to the area of emotion in speech recognition. For the remainder of the work in this thesis, this corpus provided the foundation for successive investigations into the perception of emotion from vocal acoustics. As a result of the review chapters, several research questions were formed that define the scope of this thesis. These are:

RQ1: What are the practical prerequisites for carrying out large-scale listening tests?

RQ2: Can listeners adequately capture variation of Activation and Evaluation of emotion in naturalistic speech?

RQ3: Can mood induction procedures provide naturalistic speech with sufficiently discernible levels of emotion?

RQ4: Does nonverbal naturalistic speech convey Activity and Evaluation levels that are recognisable to listeners?

RQ5: How do ratings from two perceptually different conditions (verbal and nonverbal speech) compare?

The first two questions (RQ1 and RQ2) were formed in chapter 3 after examining the various approaches to labelling emotion in speech. Having acquired a naturalistic emotional speech dataset to annotate, the first aim was to develop a tool that is accessible to a large group of people. To do this, the review in chapter 3 considered various assessment techniques and emotion representations. It had been decided to provide effect-type labels (representing observed emotions) rather than cause-type (representing internal realisation of emotions) for the given speech dataset, as these offer better validation methods (sections 3.2.3 and 3.2.5), and reflect human perception of emotion in speech. For this, perception tests (of the behavioural component) are a necessary means for assessment. The assignment of judges was also considered, and

it was determined to focus on participant magnitude and generality—including naïve listeners as opposed to solely expert' listeners who are familiar with emotion theoretical knowledge (section 3.2.6). It was established that the most suitable descriptive scheme for the annotation of the given speech dataset (characterised as naturalistic) was the dimensional approach. This approach is a more objective method and avoids issues with selecting appropriate discrete categories, which has been proven to be particularly complex for naturalistic data. Having reviewed the various dimensional models, the Activation and Evaluation dimensions are the most prevalent which, for naturalistic speech data, is commonly incorporated within frameworks that have various qualitative coding strategies (section 3.3.4). Despite the fact that the dimensional approach has been successfully implemented in numerous studies, it remains to be determined whether it is suitable for all speech datasets, such as those composed of mood inducing procedures. Thus, by adopting this framework for the online rating tool developed for this thesis, the investigations contribute to the knowledge on the practical validity of using dimensions to convey observable emotions in naturalistic speech.

Research question one (RQ1) was answered in Chapter 6 and Chapter 7. Chapter 6 documented the development of an online rating tool. The practical issues that were considered for carrying out large-scale listening tests were presented. These included a user-centered design for the tool, which ensured it was easy to use for laymen. It was designed to be suitable for repeated use to accumulate continual ratings; great emphasis was placed on the participant's overall understanding of each scale by including instructions and carrying out a survey on 7 non-expert individuals to ensure sufficient understand of the instructions. The feedback obtained from the survey suggested that participants could adequately understand the concept of Activation and Evaluation dimensions. Chapter 7 documented some of the practical difficulties of accumulating large-scale ratings. To decrease the likelihood of spurious ratings caused by fatigue and/or boredom effects, the required number of clips to be rated in a session was kept to a minimum. In spite of several reminders, it was difficult to achieve consistent daily rating from individual participants. The process of accumulating ratings, therefore, needed to be ongoing. Unfortunately, this turned out to be more time consuming than originally envisaged. The case study

carried out in Chapter 7 aimed to answer research question two (RQ2). To some degree, the low proportion of clips not rated (DNR rating) suggests that there was little confusion with the task. The inter-rater scores obtained were low. However, this is in line with many other studies that suggest Activation measures are more reliable than Evaluation measures. It is generally acknowledged that the task of rating underlying emotions in naturalistic speech is naturally difficult. Agreement levels for clips in the Active class were the highest, suggesting participants are able to adequately capture variation in Activation. For Evaluation, however, the extreme classes received low agreement, suggesting that capturing variation in extreme classes became less reliable.

Research questions three, four and five (RQ3, RQ4, and RQ5) were based on the review in Chapter 4. This chapter indicated that simulated expression does not reflect spontaneous expression, and that there is, therefore, and increasing demand for naturalistic data. While there is a large amount of control over the recording environments with acted speech, this is difficult to achieve with truly natural speech. It was argued that mood inducing procedures deliver a suitable compromise between the limitations that come with natural and acted types of materials. Research question three (RQ3) is concerned with whether there is a sufficient amount of emotional content present in the provided speech material, specified by mood induction procedures. Furthermore, this chapter considered several aspects in the perception process, which ultimately determines the obtained labels. It was recognised that in many cases it is not specified whether labels are provided based on prosodic or semantic content, or both (section 4.2). It was suggested that if the provided labels are not based exclusively on acoustic information, one should question the reliability of the relationships of acoustic variables with those labels. However, it is difficult to isolate the acoustic or the semantic content to provide labels solely on such information. To address this, inference studies that mask cues attempt to address this by removing the verbal content (section 4.1.4). It was decided to use this method to further investigate the given material and determine whether listeners would still be able to infer Activation and Evaluation levels in non-verbal naturalistic speech, thus forming research question four (RQ4). Low-pass filtering was the method chosen to remove certain acoustic cues in order

to remove the verbal content. Research question five (RQ5) is concerned with how ratings from nonverbal and verbal speech compare, by using low-pass filtering methods.

Research question three (RQ3) was answered in Chapter 7. Because emotions generally occur in natural speech, it is difficult to determine the efficacy of the MIPs by analysing the ratings from the dataset as a whole. Therefore, it was decided to categorise the clips according to which phase of the experiment they were extracted. Clips from the beginning of the experiment, prior to MIP manipulation, were assigned to one condition, and clips taken from the end were assigned to the alternative condition. It was hypothesised that the participants would be less emotionally involved at the beginning stage, prior to experimental manipulations. The results revealed a significant increase in the level of Activation for clips extracted towards the end of the experiment. However, there was no significant difference for the Evaluation scale. It suggests that mood inducing procedures were successful in inducing emotion, in terms of Activation at least. The ratings obtained for the clips as a whole exhibited a sufficient spread over all categories, indicating that mood inducing procedures provided sufficient inferable emotional content.

Chapter 8 aimed to answer question four and five (RQ4 and RQ5). For this experiment, speech was filtered and compared to its original counterpart. The results obtained from this experiment demonstrated that listeners were able to infer variation on the Activation and Evaluation dimensions in the filtered conditions as determined from the spread of the ratings, thus answering research question four (RQ4). In relation to research question five (RQ5), it appears that there is a difference between the perception of Activation and Evaluation in non-verbal and verbal speech, but that the effect is relatively small. The results showed that there was a strong correlation between agreement measures between both conditions, suggesting that acoustic variables contribute to the difficulty of reliably perceiving emotion in speech. There was a marginal difference between the reliability and agreement measures between both conditions, indicating that the task of rating subtle emotions is difficult, regardless of semantic content. It was difficult to determine if rating nonverbal speech was more reliable for Evaluation. There

was a slight increase in inter-rater reliability for the nonverbal condition, yet a slight decrease in agreement for the nonverbal condition, when considering it at the individual clip level.

9.2 Contributions of the Thesis

This thesis has made three novel contributions to the field of naturalistic emotional speech:

1. *A web-based rating tool* was developed and pilot tested during the course of this research. The design of the tool complies with the practicality of crowdsourcing, optimising labelling accuracy, minimising subjective workload, ensuring adequate accessibility of the emotion concepts, and encouraging participation. The rating tool annotates emotional speech clips on discretised Activation and Evaluation scales, with the option to not rate. Two backend databases were used, one to store participant demographics and the other to store emotional ratings of Activation and Evaluation. Additionally, an adapted design of the tool was created to allow for within-subject perception experiments. This design allows for two different emotional auditory stimuli to be compared and analysed, and to be, in the same way, administered to the online community. The adapted tool can be used as a platform to utilise different stimuli conditions in future work.

2. *A benchmark naturalistic emotional speech corpus*[1] was delivered and hosted online to be freely available to the general research community. As mentioned, part of the work in this thesis contributes to the construction of a naturalistic emotional speech corpus. The previous investigations provided high-quality emotional speech material [159, 373, 364], and gave us a solid foundation for research on labelling its emotional content. This work extends on the corpus development by providing an annotation protocol and annotation validation procedures for the acquired mood induced speech material. In this thesis, large-scale listening tests were carried out to obtain emotional dimensional ratings. The complete corpus is available to the general research community, which supplies high-quality speech samples (default download at 44.Khz/24-bit but available

[1]The **D**ublin **I**nstitute of **T**echnology, **I**nduced **E**motional (**DIT.IE**) speech corpus

up to 192Khz/24-bit), the full set of emotional ratings (Activation and Evaluation), and Ground Truth values for each clip.

3. *An analysis of the perception of non-verbal speech* was provided by comparing it to its verbal counterpart. To provide non-verbal stimuli, three different filtering conditions were assessed. Two preliminary surveys were carried out to (1) determine the most suitable filtering condition, and (2) to ensure all spoken dialogue was unintelligible. As a result, a unique filtering condition was determined. The investigation determined that there was a relatively small difference in the ratings, but a strong correlation between agreement levels in both conditions, suggesting that the difficulty in rating speech, and ambiguity of emotional content, can be attributable to the acoustics and not necessarily to the incongruence of linguistic and acoustic channels.

This thesis has considered the practical viability of annotating speech via the web and considered methods to validate annotation. Furthermore, The implementation of a controlled experimentation on speech stimuli of different conditions allowed us to determine label quality and validity. Moreover, it has broadened our knowledge regarding which voice cues are utilised to make inferences of emotion in naturalistic speech recognition.

9.3 Future Work

There are several potential directions that have arisen from the work in this thesis from which further research can be continued. Three major areas are presented here to determine how further development may proceed.

9.3.1 Acoustic Analysis

A suitable starting point for subsequent research based on this thesis is to examine the acoustic correlates associated with the annotations obtained from (1) the case study in Chapter 7, and from (2) the experiment that compared verbal and non-verbal speech in Chapter 8. Encoding

studies related to the former approach (1) should try to establish the association between the measurable acoustic parameters of the speech signal with labels that have been derived from perception tests based on the complete speech signal, comprised of prosodic and semantic content. For the latter approach (2), correlation coefficients can be compared between each condition (verbal and non-verbal) to determine whether it is adequate to correlate acoustic features with labels derived from prosodic and semantic content, or whether labels need to be derived from prosodic content alone.

The precise nature of filtering effects on the perception of speech remains to be established in both speech intelligibility, and affect in speech studies. It is evident from this thesis that the lower frequencies in naturalistic mood induced speech contain a significant amount of paralinguistic features that convey Activation and Evaluation levels. The extent to which the remaining acoustic features contribute to, or signify Activation and Evaluation remains to be investigated. It is suggested that many voice quality cues are inaudible after low-pass filtering [454, 144], yet some laryngeal voice qualities related to spectral balance can still be perceived. Because low-pass filtering affects or eliminates perceived loudness, articulation information, activity level, and formant frequencies, it would be of interest to analyse the remaining acoustic cues that listeners utilise to make inferences about Activation and Evaluation levels.

Acoustic analysis has been made more accessible in recent times by the development of several advanced software packages. One of the most widely used software packages for speech analysis is PRAAT software [449]. This software was used in the previous chapter to low-pass filter the speech clips. PRAAT is a flexible software application that is capable of analysing, manipulating, and synthesising speech. It can provide analysis of a wide range of voice cues relevant to the vocal communication of emotion, specifically those that remain after low-pass filtering has been applied. With more advanced measurement systems, researchers are faced with large acoustic feature sets, as illustrated in the INTERSPEECH challenges (section 5.8). By performing inference studies, and masking particular cues as illustrated in this thesis, the significance of particular features can be determined, and, thus, reducing potential feature sets.

Due to time restriction, acoustic analysis was limited and, therefore, not discussed in this thesis. This will be the focus of future work.

9.3.2 Cue Manipulation: Altering Speech Rhythm

The work presented in Chapter 8 sought to investigate the most prominent voice cues that listeners use to infer emotion in speech. By manipulating or removing certain cues one can investigate their roles (see section 4.1.4). Furthermore, it was mentioned that rhythm in speech is an important prosodic feature that conveys emotion (sections 5.2 and 5.4). Speech rhythm, however, is not easily defined. According to Werner and Keller [384], neither local or global rate modifications are entirely linear, and therefore, accelerating or decelerating normal speech by a fixed rate will not produce natural-sounding speech, as can be imagined. We can, however, use these artefacts to investigate their effect on the perception of emotion in naturalistic speech. To our knowledge, there are no existing emotion-in-speech studies that manipulate rhythm of speech in such a way that preserves—or at least keep the changes to a minimum—temporal or sequential organisation of pitch patterns. The reason for this is perhaps obvious to some—pitch patterns are in fact part of the rhythmic structure. As illustrated in the example in section 8.1 (see Figure 8.1), features are generally integrated, and altering one set of features can, and in most cases will, affect the other feature sets. However, just as Scherer envisaged [12], contemporary technological advances allow for more refined and natural sounding alterations in speech cues such as speech rhythm, that are relevant to emotional communication. For example, contemporary software allows for time manipulation processing such as audio stretching, often referred to *elastic audio*, without affecting the pitch—unachievable in earlier times when audiotapes were used. The elements of rhythm in speech can be systematically manipulated to attenuate features such as prominence, pauses, intensity peaks, etc. Moreover, with contemporary technology it is more accessible to alter local rate changes, such as stressing, and global rate changes, such as speech rate. With the benefit of such tools, effective rhythmic manipulations can be achieved, while minimising audible artefacts and changes in the temporal

order of feature sets such as intensity, pitch change, pitch variation, pitch contours, etc. As mentioned above (section 9.1), the tool developed for the experimental work presented in this thesis can be used as a platform to utilise different stimuli conditions. In other words, stimuli that contain rhythmic manipulations can be uploaded to the tool and effectively comparing it to non-manipulated speech and implementing the same methodology.

For this work, several areas would need to be considered. These may include areas such as the temporal perception of speech, temporal gestalt perception, rhythm and speech rate, speech stress and prominence, etc. Some of the following questions are examples that could be considered:

- How can we conceptualise speech rhythm?
- How do we measure and transcribe speech rhythm?
- What aspects of speech rhythm portray emotion?
- Is there a relationship between temporal music theory and temporal speech theory?

Providing an adequate methodology for transcribing rhythm in speech is essential for speech manipulation and analysis. For this, software based prosody analysis tools exist. For example, 'the Prosogram' transcribes prosody of speech based on intonation[2]. This system is implemented as a script in the phonetic analysis software, PRAAT[3].

9.3.3 Emotional Speech Stimuli

There are several methods to induce emotion in participants. The importance of audio of high quality was argued for here, recorded material consisting of a high-resolution signal with minimal unwanted noise. To meet these conditions a laboratory with a controlled recording environment is required. Optimal recording environments should make use of soundproof booths.

[2]More information about Prosogram can be found at: http://bach.arts.kuleuven.be/pmertens/prosogram/
[3]More information about PRAAT can be found at: http://www.fon.hum.uva.nl/praat/

A range of laboratory methods exist to induce temporary mood states as suggested in section 4.3.3. In order to further expand on the variety and size of the current corpus, future work will consider other MIPs to determine if a difference in emotional ratings can be measured between different MIP based experiments. There are, however, many restrictions with MIPs that are ethically questionable. Therefore, in order to investigate this, creative and carefully planned procedures to elicit emotion need to be devised. Some previous MIPs in contemporary research would not be suitable, for example, the use of drugs to elicit emotion in participants (Drug MIPs). Furthermore, some critical literature exists to question the validity of MIPs (e.g. [455]), hence other alternatives need to be assessed. While much of the literature acknowledges that annotating natural speech is a difficult task, it is suggested here that there is a need to further explore other MIPs, or alternative elicitation methods. Research in this area should consider existing and alternative social psychological experiments for which the following questions might be considered:

- Do ethical issues prevent us from eliciting negative emotion? If so, is it necessary to include negative evaluation?
- Do ethical issues prevent us from eliciting high activation in participants?
- Can interview based experiments elicit emotion? If so, can its authenticity and/or strength be evaluated and compared?
- Can emotion be induced to be more relevant for dimensional models?
- Can specific emotions be induced that apply to specific applications, such as *irateness* observed in call centres?
- How do we prevent participant awareness biases (e.g. *demand effects*)

Moreover, as mentioned in Chapter 8, both linguistic and acoustic cues contribute to emotional speech communication. However, both channels do not necessarily align with one another. Acoustic cues can serve to elaborate on or conflict with the conveyed meaning of the linguistic channel, for example, if one says “I am really happy about that” with an angry tone of voice (i.e.

sarcasm). The results from the experiment showed that there was strong correlation between agreement measurements of both conditions, on both scales, suggesting that the low agreement for each speech clip could be attributable to the acoustic cues. To investigate this concept further, it would be of interest to systematically select speech material so that the prosodic and semantic content are not congruent in meaning. The selection of speech stimuli that were chosen for the latter experiment (section 8.3.2) was based on the label's statistical reliability, i.e. high agreement levels that were obtained from the earlier case study. Alternatively, one could investigate a different set of speech stimuli chosen based on low agreement levels, rather than high agreement. Having inspected the material used for the investigations in this thesis, a subjective assessment suggests that an insufficient degree of incongruence exists in the speech material. In fact, to obtain such material may not be straightforward with naturalistic speech, and one may, for this reason, consider using acted speech in order to systematically generate speech with opposing paralinguistic and linguistic meaning, such as speech with sarcastic intonation. Alternatively, it may be possible to obtain such speech using sophisticated inducing procedures.

9.4 Overall Conclusions

This thesis was undertaken in relation to the following statement:

Many emotion recognition systems emulate the processes of human inferences. To recognise emotion automatically from paralinguistic information, one requires a comprehensive understanding as to the nature of the inference process of emotion from vocal acoustics—irrespective of the intertwining semantic content. Numerous acoustic features indicate emotion in speech, but the extent to which each influence perception of emotion in natural, spontaneous, mood induced speech remains unknown. Arguably, by removing, masking, or manipulating verbal/vocal cues in expressive speech, a listener's perception of expressed emotions should be constricted or misrepresented, therefore, allowing us to quantify its effects.

The acquisition of suitable naturalistic emotional speech material was first considered. Mood inducing procedures were identified as a promising compromise between acted and truly naturalistic speech data, and it was determined that dimensional ratings obtained on a large-scale basis were the appropriate means for labelling this type of speech material. This thesis describes the design, testing and evaluation of an interactive web-based listening tool developed to obtain ratings on a large-scale basis for labelling naturalistic emotional speech. Accordingly, a case study was carried out to obtain ratings for annotating and analysing the emotional content in naturalistic speech, providing the research community with a viable naturalistic emotional speech corpus. Based on this corpus, an experiment was carried out to compare the perception of verbal against nonverbal speech clips. This work identified suitable filtering conditions for non-verbal speech analysis, and determined that rating agreement levels can be attributable to acoustic (low frequency) content independently. In summary, this thesis was defended by answering five research questions in this regard (section 9.1).

Although this work has shown that low-pass filtering is a useful tool to mask semantic content, a broader selection of naturalistic speech material needs to be investigated to support the findings in this thesis, and other cue manipulation techniques need to be carried out to investigate further which voice cues contribute to reliable perceptions of emotion.

Bibliography

[1] P Shaver, J Schwartz, D Kirson, and C O'Connor. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–86, June 1987. xiii, 32, 33, 36, 42, 44

[2] H. Schlosberg. Three dimensions of emotion. *Psychological Review*, 61(2):81–88, 1954. xiii, 38

[3] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980. xiii, 39, 44

[4] R. Plutchik. *Emotions and life: perspectives from psychology, biology, and evolution*. American Psychological Association, 2003. xiii, 40

[5] TL Gehm and K. R. Scherer. Factors determining the dimensions of subjective emotional space. In K. R. Scherer, editor, *Facets of emotion: Recent research*, chapter 5, pages 99–114. Psychology Press, 1988. xiii, 41, 44

[6] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, 1988. xiii, 13, 35, 45, 46

[7] K. R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, 2005. xiii, 2, 13, 15, 16, 17, 19, 25, 26, 35, 44, 47, 48, 51, 60

[8] P J Lang. Behavioral treatment and bio-behavioral assessment: Computer applications. In Joseph B Sidowski, James H Johnson, and Thomas A Williams, editors, *Technology in Mental Health Care Delivery Systems*, pages 119–137. Ablex Pub. Corp., 1980. xiii, 52, 53

[9] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M Schröder. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA Workshop on Speech and Emotion*, Newcastle, United Kingdom, 2000. Citeseer. xiv, 53, 54, 121, 122, 143

[10] S. Steidl, A. Batliner, Dino Seppi, and Katholieke Universiteit. The Hinterland of Emotions: Facing the Open-Microphone Challenge. In *Proceedings of Affective Computing and Intelligent Interaction (ACII)*, pages 690–674, Amsterdam, 2009. xiv, 29, 55, 99, 113, 114, 116

[11] Veronique Tran. *The influence of emotions on decision-making process management teams*. PhD thesis, Université de Genève, 2004. xiv, 56

[12] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003. xiv, 27, 61, 69, 71, 73, 75, 76, 80, 88, 89, 90, 91, 92, 94, 97, 100, 102, 114, 196

[13] Jo-Anne Bachorowski. Vocal Expression and Perception of Emotion. *Current Directions in Psychological Science*, 8(2):53–57, April 1999. xiv, 25, 97, 100, 101, 102

[14] Karl Nordstrom and Peter F. Driessen. Variable pre-emphasis LPC for modeling vocal effort in the singing voice. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, pages 157–160, Montreal, Canada, 2006. xiv, 102, 114

[15] M. J. Owren and Jo-Anne Bachorowski. Measuring emotion-related vocal acoustics. *Handbook of emotion elicitation and assessment*, pages 239–266, 2007. xv, 91, 94, 100, 107, 108, 109, 113, 114, 115, 116, 117

[16] R. Cowie and R. R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32, 2003. xviii, 15, 19, 25, 29, 30, 31, 33, 34, 37, 63, 138

[17] Klaus Krippendorff. Computing Krippendorff s Alpha-Reliability. *Departmental Papers (ASC)*, 2007. xviii, 144, 145, 148, 171, 172

[18] K. R. Scherer and Jeffery Pittam. Vocal Expression and Communication. In *Handbook of Emotions*, chapter 13, pages 185–195. The Guilford Press, New York, London, 1993. 1

[19] N Fragopanagos and J G Taylor. Emotion recognition in human–computer interaction. *Neural networks*, 18(4):389–405, May 2005. 2, 70, 90, 93, 104

[20] Louis ten Bosch. Emotions, speech and the ASR framework. *Speech Communication*, 40:213–225, 2003. 2

[21] M Schröder, R. Cowie, E. Douglas-Cowie, M Westerdijk, and Stan Gielen. Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. In *Proceedings of Eurospeech*, volume 1, pages 87–90, Aalborg, Denmark, 2001. 2

[22] C. Becker-Asano and H. Ishiguro. Laughter in Social Robotics–no laughing matter. In *Intl. Workshop on Social Intelligence Design (SID2009)*, pages 287–300, 2009. 2, 56

[23] C. Busso, Zhigang Deng, M. Grimm, Ulrich Neumann, and S. S. Narayanan. Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1075–1086, March 2007. 2, 52, 69

[24] L. Devillers, L. Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005. 2, 42, 63

[25] L Feldman Barrett. Constructing Emotion. *Psychological Topics*, 3:856–860, 2011. 2, 6

[26] R. Cowie, N. Sussman, and A. Ben-Ze'ev. Emotion: Concepts and Definitions. In P. Petta, C. Pelachaud, and R. Cowie, editors, *Emotion-Oriented Systems: The Humaine Handbook*, pages 9–30. Springer Berlin Heidelberg, 2011. 2, 28

[27] R. R. Cornelius. Theoretical Approaches to emotion. In M. S. Clark, editor, *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 3–10, Newcastle, Northern Ireland, 2000. 2, 7, 11, 12, 13, 15, 32

[28] Petri Laukka, Daniel Neiberg, Mimmi Forsell, Inger Karlsson, and Kjell Elenius. Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language*, 25(1):84–104, January 2011. 2, 108

[29] P. Juslin and K. R. Scherer. Vocal Expression of Affect. In *The new handbook of methods in nonverbal behavior research*, chapter 3, pages 65–135. Oxford University Press, Oxford, series in edition, 2005. 2, 78, 79, 86, 90, 91, 92, 93, 94, 97

[30] R. Cowie, Cate Cox, Jean-Claude Martin, A. Batliner, Dirk Heylen, and Kostas Karpouzis. Issues in Data Labelling. In Roddy Cowie, Catherine Pelachaud, and Paolo Petta, editors, *Emotion-Oriented Systems: The Humaine Handbook*, pages 213–241. Springer Berlin Heidelberg, 2011. 6, 30, 37, 50, 51, 58, 65, 107, 121, 137

[31] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, and K. Elenius. Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language*, April 2010. 6, 29, 76, 77, 92, 115

[32] M Schröder. Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In *Workshop on Affective Dialogue Systems*, pages 209–220, Kloster Irsee, Germany, 2004. Springer. 6, 43, 63, 152, 184

[33] C. Darwin. *The Expression of the Emotions in Man and Animals*. New York: Oxford University Press (Original work published 1872), 1998. 7, 8, 70

[34] R. Plutchik. *Emotion: A psychoevolutionary synthesis*. Harper and Row, 1980. 7, 40, 44, 53

[35] G. A. Bryant and H.C. Barrett. Vocal Emotion Recognition Across Disparate Cultures. *Journal of Cognition and Culture*, 8(1):135–148, April 2008. 7, 8, 23, 25, 139

[36] P. Ekman. Strong evidence for universals in facial expression: A reply to russell's mistaken critique. *Psychological Bulletin*, 115:268–287, 1994. 7, 30, 31

[37] J. Panksepp. Mood changes. *Handbook of clinical neurology*, 1:271–285, 1985. 7

[38] Jaak Panksepp. On the embodied neural nature of core emotional affects. *Journal of Consciousness Studies, 12*, 8(10):158–184, 2005. 7

[39] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, K. R. Scherer, M. Tomita, and A. Tzavaras. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53:712–717, 1988. 8

[40] P. Ekman, W.V. Friesen, and S. Ancoli. Facial Signs of Emotional Experience. *Journal of Personality and Social Psychology*, 39(6):1125–1134, 1980. 8, 24

[41] Randolph R. Cornelius. *The science of emotion. Research and tradition in the psychology of emotions.* Upper Saddle River (NJ): Prentice-Hall, 1996. 8, 15

[42] Jo-Anne Bachorowski and Michael J Owren. *Handbook of emotions*, chapter 12: Vocal expressions of emotion, pages 196–210. The Guilford Press, 3rd edition, 2008. 8, 97, 102, 117

[43] K. R. Scherer, R. Banse, and H. G. Wallbott. Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. *Journal of Cross-Cultural Psychology*, 32(1):76–92, January 2001. 8, 75, 139

[44] Robert C. Solomon. The philosophy of emotions. In M. Lewis, J.M. Haviland-Jones, and Lisa Feldman Barrett, editors, *Handbook of emotions*, chapter 1, pages 3–16. The Guilford Press, 2008. 8

[45] R. Descartes. *On the passions of the soul*. (Voss, S. H., trans.). Hackett, Indianapolis, (Original work published in 1649), 1989. 9

[46] D. Hume. A treatise of human nature. *Journal of Philosophy*, 73:733–757, 1888. 9

[47] W. James. What is an emotion? *Mind*, 9:188–205, 1884. 9

[48] R. Buck. Nonverbal behaviour and the theory of emotion: the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 38:811–824, 1980. 9

[49] J. N. Capella. The facial feedback hypothesis in human interaction. *Journal of Language and Social Psychology*, 12:13–29, 1993. 9

[50] J.T. Lanzetta, J Cartwright-Smith, and R. E. Kleck. Effects of nonverbal dissimulation on emotional experience and autonomic arousal. *Journal of Personality and Social Psychology*, 33(3):354–370, 1976. 9

[51] F. Strack, L. L. Martin, and S. Stepper. Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54:768–777, 1988. 9, 92

[52] J. T. Cacioppo and R. E. Petty. Electromyograms as measures of extent and affectivity of information processing. *American Psychologist*, 36:441–456, 1981. 9

[53] J. T. Cacioppo, R. E. Petty, M. E. Losch, and H. S. Kim. Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology*, 50(2):260–268, 1986. 9

[54] R. Neumann, M. Hess, S. Schulz, and G. W. Alpers. Automatic behavioural responses to valence: Evidence that facial action is facilitated by evaluative processing. *Cognition and emotion*, 19(4):499–513, 2005. 9

[55] L. F. Barrett, K. S. Quigley, E. Bliss-Moreau, and K. R. Aronson. Interoceptive sensitivity and self-reports of emotional experience. *Journal of Personality and Social Psychology*, 87(5):684–697, 2004. 9

[56] W. B. Cannon. The james-lange theory of emotion: A critical examination and an alternative theory. *American Journal of Psychology*, 39:10–124, 1927. 10

[57] Stefan Wiens. Interoception in emotional experience. *Current Opinion in Neurology*, 18(4):442–7, August 2005. 10

[58] Pilar Cobos, María Sánchez, Carmen García, María Nieves Vera, and Jaime Vila. Revisiting the James versus Cannon debate on emotion: startle and autonomic modulation in patients with spinal cord injuries. *Biological Psychology*, 61(3):251–69, November 2002. 10

[59] K. Chwalisz, E. Diener, and D Gallagher. Autonomic arousal feedback and emotional experience: Evidence from the spinal cord injured. *Journal of Personality and Social Psychology*, 54:820–28, 1988. 10

[60] J. W. Papez. A proposed mechanism of emotion. *Archives of Neurological Psychiatry*, 38:725–743, 1937. 10

[61] P. D. MacLean. Psychosomatic disease and the "visceral brain": Recent developments bearing on the papez theory of emotion. *Psychosomatic Medicine*, 11:338–353, 1949. 10

[62] Tim Dalgleish and M.J. Power. *Handbook of cognition and emotion*. John Wiley & Sons, Ltd., 1999. 11

[63] S. Schachter and J. E. Singer. Cognitive, social and physiological determinants of emotional state. *Psychological Review*, 69(5):379–399, 1962. 11

[64] A. R. Damasio. Descartes' error: Emotion, reason, and the human brain. *New York: Putnam*, 1994. 11

[65] L.F. Barrett. Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1):28, March 2006. 11, 14, 75

[66] L.F. Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20, 2006. 11

[67] Tim Dalgleish. The emotional brain. *Nature Reviews Neuroscience*, 5:583–9, July 2004. 11, 20

[68] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172, 2003. 11, 14, 39

[69] M. B. Arnold. *Emotion and personality*. Columbia University Press, New York, 1960. 12

[70] N H Frijda. *The Emotions*. Studies in Emotion and Social Interaction. Cambridge University Press, 1986. 12, 15, 16, 22, 43

[71] K. R. Scherer. On the Nature and Function of Emotion: A Component Process Approach. In *Approaches to emotion*, chapter 14, pages 293–317. NJ: Erlbaum, Hillsdale, 1984. 12, 35, 41

[72] K. R. Scherer. Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, and T. Johnstone, editors, *Appraisal processes in emotion: Theory, Methods, Research*, pages 92–120. Oxford University Press, New York and Oxford, 2001. 12

[73] K. R. Scherer. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99(2):143–165, March 1986. 13, 25, 69, 70, 71, 88, 100, 111, 112

[74] A. Ortony, D. A. Norman, and W. Revelle. *Who needs emotions: The brain meets the machine*, chapter 7, Affect and proto-affect in effective functioning, pages 174–202. Oxford University Press, New York, 2005. 13, 45

[75] R Harre. *The Social Construction of Emotions*. Blackwell Pub, Oxford, England, 1986. 13

[76] J. R. Averill. A constructivist view of emotion. In R. Plutchik and H. Kellerman (Eds.), editors, *Emotion: Theory, research and experience*, volume 1, pages 305–339. New York: Academic Press, 1980. 13, 14, 15

[77] J. Prinz. Which emotions are basic. In D. Evans and P. Cruse, editors, *Emotion, evolution, and rationality*, chapter 4, pages 69–87. Oxford University Press, 2004. 14, 32

[78] J R Averill. *Anger and aggression: An essay on emotion.* New York: Springer-Verlag, 1982. 14

[79] P. N. Stearns. History of emotions. In M. Lewis, J.M. Haviland-Jones, and L.F. Barrett, editors, *Handbook of Emotions*, chapter 2, pages 17–31. The Guilford Press, New York, London, 3rd edition, 2008. 14

[80] P Ekman. Universals and cultural differences in facial expressions of emotion. *University of Nebraska*, pages 207–283, 1972. 14

[81] J. Panksepp. The periconscious substrates of consciousness: Affective states and the evolutionary origins of self. *Journal of Consciousness Studies*, 5:566–582, 1998. 14

[82] J. Panksepp. *Handbook of Emotions*, chapter 4: The Affective Brain and Core Consciousness, pages 47– 67. The Guilford Press, 3rd edition, 2008. 14

[83] David Matsumoto, Bob Willingham, and Andres Olide. Sequential dynamics of culturally moderated facial expressions of emotion. *Psychological science*, 20(10):1269–75, October 2009. 14

[84] P. Ekman. Darwin and cross cultural studies of facial expression. *P. Ekman (Ed.), Darwin and facial expression: A century of research in review*, 1973. 15

[85] P Ekman and W Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978. 15, 24, 50

[86] C. E. Izard. *Human Emotions*. Emotions, Personality, and Psychotherapy. Springer, 1977. 15, 24, 31, 32

[87] M Schröder. *Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis.* PhD thesis, Universitat des Saarlandes: 288, 2004. 15, 27, 39, 45, 70, 72, 73, 77, 112

[88] IB Mauss and MD Robinson. Measures of emotion: A review. *Cognition and emotion*, 23(2):1–23, 2009. 15, 16, 22, 24, 26

[89] Michael D. Robinson and Gerald L. Clore. Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128(6):934–960, 2002. 15

[90] P J Lang, M K Greenwald, M M Bradley, and a O Hamm. Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–73, May 1993. 15, 24

[91] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3):169–200, 1992. 15, 25

[92] Iris B Mauss, Robert W Levenson, Loren McCarter, Frank H Wilhelm, and James J Gross. The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, 5(2):175–90, June 2005. 17, 26

[93] Michael D. Robinson and Gerald L. Clore. Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology*, 83(1):198–215, 2002. 19

[94] Michael D. Robinson and Lisa Feldman Barrett. Belief and Feeling in Self-reports of Emotion: Evidence for Semantic Infusion Based on Self-esteem. *Self and Identity*, 9(1):87–111, January 2010. 19

[95] J Panksepp. Neurologizing the Psychology of Affects: How Appraisal-Based Constructivism and Basic Emotion Theory Can Coexist. *Perspectives on Psychological Science*, 2(3):281–296, 2007. 20

[96] Annett Schirmer, Sonja a Kotz, and Angela D Friederici. Sex differentiates the role of emotional prosody during word processing. *Cognitive brain research*, 14(2):228–33, August 2002. 20, 168

[97] S. A. Kotz and S. Paulmann. When emotional prosody and semantics dance cheek to cheek: ERP evidence. *Brain research*, 115:107–18, June 2007. 20

[98] Guillaume Chanel, Julien Kronegg, Didier Grandjean, and Thierry Pun. Emotion Assessment: Arousal Evaluation Using EEG's and Peripheral Physiological Signals. In *Multimedia Content Representation, Classification and Security, International Workshop (MRCS 2006)*, Istanbul, Turkey, 2005. Springer LNCS. 20, 52

[99] Ingo Hertrich, Klaus Mathiak, Werner Lutzenberger, and Hermann Ackermann. Processing of dynamic aspects of speech and non-speech stimuli: a whole-head magnetoencephalography study. *Cognitive Brain Research*, 17(1):130–9, June 2003. 20

[100] Thomas R. Knösche, Sonja Lattner, Burkhard Maess, Michael Schauer, and Angela D. Friederici. Early Parallel Processing of Auditory Word and Voice Information. *NeuroImage*, 17(3):1493–1503, November 2002. 20

[101] M S George, P I Parekh, N Rosinsky, T A Ketter, T A Kimbrell, K M Heilman, P Herscovitch, and R M Post. Understanding emotional prosody activates right hemisphere regions. *Arch Neurol*, 53(7):665–670, July 1996. 20

[102] Pascal Belin, Monica Zilbovicius, Sophie Crozier, Lionel Thivard, Fontaine, Anne, Marie-Cécile Masure, and Yves Samson. Lateralization of Speech and Auditory Temporal Processing. *Journal of Cognitive Neuroscience*, 10(4):536–540, July 1998. 20

[103] D. K. Tracy, D. K Ho, Owen O'Daly, P. Michalopoulou, Lisa C Lloyd, Eleanor Dimond, Kazunori Matsumoto, and Sukhwinder S Shergill. It's not what you say but the way that you say it: an fMRI study of differential lexical and non-lexical prosodic pitch processing. *BMC neuroscience*, 12:128, January 2011. 20

[104] Rachel L C Mitchell, Rebecca Elliott, Martin Barry, Alan Cruttenden, and Peter W R Woodruff. The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia*, 41:1410–1421, 2003. 20

[105] T. Johnstone, Carien M van Reekum, Terrence R Oakes, and Richard J Davidson. The voice of emotion: an FMRI study of neural responses to angry and happy vocal expressions. *Social cognitive and affective neuroscience*, 1(3):242–9, December 2006. 20, 70

[106] AM Dale and MI Sereno. Improved localizadon of cortical activity by combining eeg and meg with mri cortical surface reconstruction: A linear approach. *Journal of Cognitive Neuroscience*, 5(2):162–176, 1993. 20

[107] Lisa Feldman Barrett. Are Emotions Natural Kinds? *Perspectives on Psychological Science*, 1(1):28–58, March 2006. 21

[108] Sylvia D Kreibig, Frank H Wilhelm, Walton T Roth, and James J Gross. Cardiovascular, electrodermal, and respiratory response patterns to fear- and sadness-inducing films. *Psychophysiology*, 44(5):787–806, September 2007. 21

[109] G Stemmler, M Heldmann, C. A. Pauls, and T Scherer. Constraints for emotion specificity in fear and anger: the context counts. *Psychophysiology*, 38(2):275–91, March 2001. 21

[110] MH Schut, Kees Tuinenbreijer, EL Broek, and JHDM Westerink. Biometrics for Emotion Detection (BED): Exploring the combination of Speech and ECG. In *Proceedings of the 1st International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications (B-Interface 2010)*, pages 56–66, Valencia, Spain, 2010. INSTICC Press. 21, 22

[111] John Hl Hansen, Wooil Kim, Mandar Rahurkar, Evan Ruzanski, and James Meyerhoff. Robust Emotional Stressed Speech Detection Using Weighted Frequency Subbands. *EURASIP Journal on Advances in Signal Processing*, 2011(1):906789, 2011. 21, 22

[112] Sylvain Delplanque, Didier Grandjean, and C Chrea. Sequential Unfolding of Novelty and Pleasantness Appraisals of Odors: Evidence From Facial Electromyography and Autonomic Reactions. *Emotion*, 9(3):316–328, 2009. 21, 22

[113] M Ilves and V Surakka. Emotions, anthropomorphism of speech synthesis, and psychophysiology. *Emotions in the Human Voice. Culture and Perception*, 2009. 21, 22

[114] Patrick Gomez and Brigitta Danuser. Relationships between musical structure and psychophysiological measures of emotion. *Emotion*, 7(2):377–387, 2007. 21

[115] N. Amir, S. Ron, and N. Laor. Analysis of an emotional speech corpus in Hebrew based on objective criteria. In *Proc. ISCA Workshop Speech and Emotion*, pages 29–33, Belfast, 2000. 21, 22, 88, 107

[116] J. T. Larsen, G. G. Berntson, K. M. Poehlmann, T. A. Ito, and J. T. Cacioppo. The psychophysiology of emotion. In M. Lewis, J.M. Haviland-Jones, and Lisa Feldman Barrett, editors, *Handbook of emotions*, chapter 11, pages 180–195. The Guilford Press, 2010. 21

[117] MARGARET M. BRADLEY and PETER J. LANG. *Emotion and motivation*. Cambridge University Press, 2007. 21, 26

[118] R. Benjamin Knapp;, Jonghwa Kim; And, and Elisabeth Andre. Physiological Signals and Their Use in Augmenting Emotion Recognition for Human-Machine Interaction. In P. Petta, C. Pelachaud, and R. Cowie, editors, *Emotion-Oriented Systems: The Humaine Handbook*, pages 133 – 159. Springer Berlin Heidelberg, 2011. 22

[119] Mirja Ilves and Veikko Surakka. Heart Rate Responses to Synthesized Affective Spoken Words. *Advances in Human-Computer Interaction*, 2012:1–6, 2012. 22

[120] K. Alter, S. A. Kotz, U. Toepel, M. Besson, A. Schirmer, and A. D. Friederici. Accentuation and emotions-Two different systems? *Speech and Emotion*, 2000. 22

[121] Tom Johnstone and K. R. Scherer. The effects of emotions on voice quality. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, pages 2029– 2032, San Francisco, USA, 1999. 22, 111

[122] Jonghwa Kim, E André, Matthias Rehm, Thurid Vogt, and Johannes Wagner. Integrating information from speech and physiological signals to achieve emotional sensitivity. *INTERSPEECH*, 2005. 22

[123] M Ilves and V Surakka. Subjective and physiological responses to emotional content of synthesized speech. *Computer Animation and Social Agents (CASA)*, 2004. 22

[124] P J Lang, M M Bradley, and B N Cuthbert. Emotion, attention, and the startle reflex. *Psychological review*, 97(3):377–95, July 1990. 22

[125] F. Ramus. Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, 2(1):85–115, October 2002. 23

[126] Elizabeth S Paul, Emma J Harding, and Michael Mendl. Measuring emotional processes in animals: the utility of a cognitive approach. *Neuroscience and biobehavioral reviews*, 29(3):469–91, May 2005. 23

[127] Gary McKeown, William Curran, Ciaran McLoughlin, Harry J. Griffin, and Nadia Bianchi-Berthouze. Laughter induction techniques suitable for generating motion capture data of laughter associated body movements. *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–5, April 2013. 23

[128] Mark Coulson. Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, 2004. 23

[129] J. A. Harrigan. Proxemics, kinesics and gaze. In *The new handbook of Methods in Nonverbal Behavior Research*, chapter 4, pages 137 – 198. Oxford University Press, Oxford, 2005. 23

[130] Mark Chen. Immediate Behavioral Predispositions to Approach or Avoid the Stimulus. *Personality and Social Psychology Bulletin*, 25(2):215–224, 1999. 23

[131] Susan Bamford and Robert Ward. Predispositions to approach and avoid are contextually sensitive and goal dependent. *Emotion (Washington, D.C.)*, 8(2):174–83, May 2008. 23

[132] Nicole L Nelson and J. A. Russell. Children's understanding of nonverbal expressions of pride. *Journal of experimental child psychology*, 111(3):379–85, March 2012. 23

[133] Sabine Stepper and Fritz Strack. Proprioceptive determinants of emotional and nonemotional feelings. *Journal of Personality and Social Psychology*, 64(2):211, 1993. 24

[134] Holger Hoffmann, Harald C Traue, Franziska Bachmayr, and Henrik Kessler. Perception of Dynamic Facial Expressions of Emotion. *Perception of Dynamic Facial Expressions of Emotion*, pages 175 – 178, 2006. 24

[135] S S Tomkins. *Affect Imagery Consciousness: The complete edition*. Springer, New York, 2008. 24, 31

[136] P. Ekman and E L Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Series in Affective Science. Oxford University Press, USA, 2005. 24

[137] JF Cohn and P. Ekman. Measuring facial action. In *The new handbook of Methods in Nonverbal Behavior Research*, chapter 2, pages 9–64. Oxford University Press, Oxford, 2005. 24

[138] Klaus Schneider and Ingrid Josephs. The expressive and communicative functions of preschool children's smiles in an achievement-situation. *Journal of Nonverbal Behavior*, 15(3):185–198, 1991. 24, 70

[139] J. A. Russell, Jo-Anne Bachorowski, and Jose-Miguel Fernandez-Dols. Facial and vocal expressions of emotion. *Annual review of psychology*, 54:329–49, January 2003. 24, 26, 97, 117

[140] EL Rosenberg and P. Ekman. Coherence Between Expressive and Experiential System in Emotion. *Cognition & Emotion*, 8(3):201–229, 1994. 24

[141] N. Amir, B.-C. Almogi, and R. Gal. Perceiving Prominence and Emotion in Speech a Cross Lingual Study. In *Proc. Speech Prosody*, pages 375–378, Nara, Japan, 2004. 25, 97, 99, 107

[142] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. *in Interspeech Lissabon, Portugal*, 2005. 25

[143] R. Rosenthal, J.A. Hall, M.R. DiMatteo, P.L. Rogers, and D. Archer. *Sensitivity to Nonverbal Communication: The PONS Test*. Johns Hopkins University Press, 1979. 25, 51

[144] H. G. Wallbott and K. R. Scherer. Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, 51(4):690–699, 1986. 25, 86, 89, 195

[145] R. van Bezooijen and L. Boves. The effects of low-pass filtering and random splicing on the perception of speech. *Journal of Psycholinguistic Research*, 15(5):403–17, September 1986. 25, 78, 86, 181

[146] M. Knoll, M. Uther, and A. Costall. Effects of low-pass filtering on the judgment of vocal affect in speech directed to infants, adults and foreigners. *Speech Communication*, 51(3):210–216, March 2009. 25, 78, 84, 85, 86, 181, 182

[147] L. Leinonen, T. Hiltunen, I. Linnankoski, and M-L. Laakso. Expression of emotionalmotivational connotations with a one-word utterance. *The Journal of the Acoustical society of America*, 102(3):1853–63, 1997. 25, 70

[148] P. Laukka, P. N. Juslin, and R. Bresin. A dimensional approach to vocal expression of emotion. *Cognition and emotion*, 19(5):633–653, 2005. 25

[149] Kornelia Gentsch, Didier Grandjean, and K. R. Scherer. Coherence explored between emotion components: Evidence from event-related potentials and facial electromyography. *Biological psychology*, December 2013. 25, 26

[150] George Bonanno and Dacher Keltner. The coherence of emotion systems: Comparing "online" measures of appraisal and facial expressions, and selfreport. *Cognition & Emotion*, 18(3):431–444, April 2004. 25

[151] K. R. Scherer. The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, 23(7):1307–1351, November 2009. 26, 63, 92

[152] M. Grimm, K. Kroschel, E. Mower, and S. S. Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787–800, 2007. 26, 42, 52, 53, 61, 62, 63, 93, 104, 116, 152

[153] Raul Fernandez and R.W. Picard. Modeling drivers' speech under stress. *Speech Communication*, 40(1-2):145–159, 2003. 27, 49

[154] R. Cowie. Describing the emotional states expressed in speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland, 2000. 27, 28, 87

[155] Richard Craggs and M Wood. A two dimensional annotation scheme for emotion in dialogue. *AAAI Spring Symposium: Exploring Attitude and Affect in Text*, 2004. 28

[156] R. Cowie, E. Douglas-Cowie, N. Tsatatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J G Taylor. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, 22(1), January 2001. 28, 31, 43, 49, 70, 90, 93, 110, 182

[157] H. Mori, T. Satake, M. Nakamura, and H. Kasuya. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, 53(1):36–50, August 2010. 28, 43, 63, 70, 76, 77, 89, 95

[158] A. Batliner, S. Steidl, Christian Hacker, and Elmar Nöth. Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech. *User Modeling and User-Adapted Interaction*, 18(1-2):175–206, October 2007. 29, 41, 42, 44, 137

[159] B. Vaughan. *Naturalistic Emotional Speech Corpora with Large Scale Emotional Dimension Ratings*. PhD thesis, Dublin Institute of Technology, 2011. 29, 63, 89, 92, 95, 96, 133, 134, 135, 137, 152, 193

[160] A. Tarasov, S.J. Delany, and C. Cullen. Using crowdsourcing for labelling emotional speech assets. In *W3C workshop on Emotion ML*, Paris, France, 2010. Dublin Institute of Technology. 29, 139

[161] Jeff Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, New York, NY, USA, 2008. 29, 139

[162] Vamshi Ambati, Stephan Vogel, and JG Carbonell. Active Learning and Crowd-Sourcing for Machine Translation. *Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 2169–2174, 2010. 29

[163] Padhraic Smyth, Usama Fayyad, and Michael Burl. Inferring ground truth from subjective labelling of venus images. *Advances in Neural Information Processing Systems*, (7):1085–1092, 1995. 29

[164] Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. *IEEE Computer Vision and Pattern Recognition Workshops (CVPR)*, pages 1–8, 2008. 29

[165] A Brew, Derek Greene, and P Cunningham. Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In *19th European Conference on Artificial Intelligence*, pages 145–150, 2010. 29

[166] PY Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing (ALNP)*, pages 27–35, 2009. 29

[167] Rion Snow, B O'Connor, Daniel Jurafsky, and AY Ng. Cheap and Fast—But is it Good? Evaluating Non-expert Annotations for Natural Language Tasks. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263, 2008. 29

[168] James R Averill. *A semantic atlas of emotional concepts*, volume 5. American Psychological Association, 1975. 30, 38

[169] Beverley Fehr and J. A. Russell. Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, 113(3):464–486, 1984. 30

[170] K. R. Scherer. Emotion as a Multicomponent Process: A model and some cross-cultural data. *Review of Personality & Social Psychology*, 5:37–63, 1984. 30, 43, 44, 46, 47

[171] Ralph B. Hupka, Alison P. Lenton, and Keith a. Hutchison. Universal development of emotion categories in natural language. *Journal of Personality and Social Psychology*, 77(2):247–278, 1999. 30

[172] Jerry R Hobbs, Andrew Gordon, and Marina Rey. The Deep Lexical Semantics of Emotions Identifying the Core Emotion Words. In K. Ahmad, editor, *Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology*, pages 27–34. Springer, 2008. 30

[173] R Plutchik. *Emotions: A general psychoevolutionary theory*, volume 1, pages 197–219. Erlbaum, Hillsdale, NJ, 1984. 30, 31

[174] K. R. Scherer. Toward a concept of modal emotions. In P. Ekman and R. (Eds.) Davidson, editors, *The Nature of Emotion: Fundamental Questions*, pages 25–31. Oxford University Press, Oxford, 1994. 30

[175] R. S. Lazarus. *The Nature of Emotion. Fundamental Questions*, chapter The stable and the unstable in emotion, pages 79–85. Oxford University Press, Oxford, 1994. 30

[176] K. R. Scherer. Appraisal Theory. In T. Dalgleish and M. J. Power, editors, *Handbook of cognition and emotion*, chapter 30, pages 637–663. Wiley, 1999. 30, 49

[177] J. A. Russell and L F Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805–19, May 1999. 30

[178] B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2011 Speaker State Challenge. In *12th Annual Conference of the International Speech Communication Association*, pages 2–5, Florence, Italy, 2011. 31, 118

[179] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, Johannes Wagner, L. Devillers, L. Vidrascu, Vered Aharonson, and Loic Kessous. Whodunnit – Searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language*, 2010. 31, 32, 90

[180] S. Steidl, A. Batliner, D. Seppi, and B. Schuller. On the Impact of Children's Emotional Speech on Acoustic and Language Models. *Journal on Audio, Speech, and Music Processing (EURASIP)*, pages 1–15, 2010. 31

[181] A. Ortony and Turner. What's Basic About Basic Emotions? *Psychological Review*, 97(3):315–331, 1990. 31

[182] Paul Ekman. Basic Emotions. In T. Dalgleish and M. Power, editor, *Handbook of Cognition and Emotion*, number 1992, chapter 3. John Wiley & Sons, Ltd., Sussex, 1999. 31, 32, 33

[183] A. S. Walker-Andrews. Infants' Perception of Expressive Behaviors: Differentiation of Multimodal Information. *Psychological bulletin*, 121(3):437–56, May 1997. 32

[184] P Ekman, W Friesen, M O'Sullivan, A Chan, I Diacoyanni-Tarlatzis, and K Heider. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53:712–717, 1987. 32

[185] P. Ekman. Facial expression and emotion. *American Psychologist*, 48:384–392, 1993. 32

[186] PR Shaver, S Wu, and JC Schwartz. Cross-cultural similarities and differences in emotion and its representation. In Margaret S. Clark, editor, *Emotion*, chapter 7, pages 175–212. Sage Publications, 1992. 32

[187] Hatice Gunes and Björn Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, July 2012. 32, 63

[188] A. Batliner, Björn Schuller, Dino Seppi, S. Steidl, L. Devillers, Laurence Vidrascu, T. Vogt, Vered Aharonson, and Noam Amir. The Automatic Recognition of Emotions in Speech. In Roddy Cowie, Catherine Pelachaud, and Paolo Petta, editors, *Emotion-Oriented Systems: The Humaine Handbook*, Cognitive Technologies, pages 71–99. Springer Berlin Heidelberg, 2011. 32, 74

[189] R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614–36, March 1996. 33, 48, 61, 69, 70, 75, 88, 89, 93, 99

[190] V Aubergé, N Audibert, and A Rilliard. Auto-annotation: an alternative method to label expressive corpora. In *LREC 2006 Workshop on Emotional Corpora*, pages 45–46, Genova, Italy, 2006. 35

[191] E. Douglas-Cowie, L. Devillers, and J.C Martin. Multimodal databases of everyday emotion: facing up to complexity. *INTERSPEECH*, pages 813–816, 2005. 35

[192] Aurélie Zara, Valérie Maffiolo, J. Martin, and L. Devillers. Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. *Affective Computing and Intelligent Interaction*, pages 464–475, 2007. 37

[193] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and Others. Combining efforts for improving automatic classification of emotional user states. In *Proc. IS-LTC*, pages 240–245, Ljubliana, 2006. 37, 64, 82, 88

[194] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J. C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, pages 488–500, Lisbon, Portugal, 2007. 37, 54, 62, 63, 91

[195] J. A. Russell, Anna Weiss, and Gerald a. Mendelsohn. Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, 57(3):493–502, 1989. 37

[196] K. R. Scherer, Elise Dan, and Anders Flykt. What determines a feeling's position in affective space? A case for appraisal. *Cognition & Emotion*, 20(1):92–113, January 2006. 37, 42, 47

[197] N. Jaworska and A. Chupetlovska-Anastasova. A Review of Multidimensional Scaling (MDS) and its Utility in Various Psychological Domains. *Applied Psychological Measurement*, 5(1):1–10, September 2009. 37

[198] Steven M Holland. Non-metric Multidimensional Scaling (MDS), 2008. 37

[199] W Wundt. *Grundzüge der physiologischen Psychologie [Fundamentals of physiological psychology]*. Wilhelm Engelmann, Leipzig, 1874. 37, 44

[200] H Scholsberg. A scale for the judgment of facial expressions. *Journal of Experimental Psychology*, 29(6):497–510, 1941. 38, 44

[201] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, April 2003. 38, 57, 62, 65, 87, 89

[202] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The measurement of meaning*. University of Illinois Press, 1957. 38, 44

[203] A. Mehrabian and J. A. Russell. *An approach to environmental psychology*. MIT Press, Cambridge, MA, 1974. 39, 44

[204] James A. Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, September 1977. 39, 52

[205] J. B. Kruskal and M. Wish. *Multidimensional scaling*. Quantitative Applications in the Social Sciences. Sage University, Beverly Hills and London, 1978. 42

[206] J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115–129, 1964. 42

[207] Martijn Goudbeek and K. R. Scherer. Beyond arousal: valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, 128(3):1322–36, September 2010. 42, 43

[208] Johnny R J Fontaine, K. R. Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–7, December 2007. 42, 43, 44, 51

[209] Charles E Osgood. Dimensionality of the Semantic Space for Communication via Facial Expressions. *Scandinavian Journal of Psychology*, 7(1):1–30, 1966. 44

[210] Lynn E Bush. Individual differences multidimensional scaling of adjectives denoting feelings. *Journal of Personality and Social Psychology*, 25(1), 1973. 44

[211] Rex S. Green and Norman Cliff. Multidimensional comparisons of structures of vocally and facially expressed emotion. *Perception & Psychophysics*, 17(5):429–438, September 1975. 44

[212] D Watson and A Tellegen. Toward a consensual structure of mood. *Psychological Bulletin*, 98(2):219–235, 1985. 44

[213] LA Feldman. Variations in the circumplex structure of mood. *Personality and Social Psychology Bulletin*, 1995. 44

[214] Timothy Church, Marcia S Katigbak, Jose Alberto S Reyes, and Stacia M Jensen. Language and Organisation of Filipino Emotion Concepts: Comparing Emotion Concepts and Dimensions across Cultures. *Cognition and Emotion*, 12(1):63–92, 1998. 44

[215] L.F. Barrett. Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus. *Cognition and Emotion*, 12(4):579–599, December 1998. 44

[216] J. C. Speisman, R. S. Lazarus, A. Mordkoff, and L. Davison. Experimental reduction of stress based on ego-defense theory. *Journal of Abnormal and Social Psychology*, 68:367–380, 1964. 45

[217] C. Becker-Asano. *WASABI: Affect simulation for agents with believable interactivity*. PhD thesis, University of Bielefeld, Faculty of Technology, 2008. 45

[218] Christoph Bartneck. Integrating the OCC model of emotions in embodied characters. In *Proceedings of the Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges*, Melbourne, Australia, 2002. Citeseer. 45

[219] David Sander, Didier Grandjean, and K. R. Scherer. A systems approach to appraisal mechanisms in emotion. *Neural networks: the Official Journal of the International Neural Network Society*, 18(4):317–52, May 2005. 46, 47

[220] J. A. Russel. Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology 45:*, 45:1281–8, 1983. 47

[221] L Devillers, R. Cowie, J-c Martin, S Abrilian, and M Mcrorie. Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. In *5th International Conference on Language Resources and Evaluation (LREC)*, pages 1105–1110, Genoa, 2006. 48, 51, 63

[222] K. R. Scherer and M.R. Zentner. Emotional effects of music: Production rules. In P. N. Juslin and J. A. Sloboda, editors, *Music and emotion: Theory and research*, chapter 16, pages 361–392. Oxford University Press, New York, 2001. 49

[223] Nathalie Lanctôt and Ursula Hess. The timing of appraisals. *Emotion*, 7(1):207–12, February 2007. 49

[224] P Friesen, W.; Ekman. EMFACS-7: Emotional Facial Action Coding System. Technical report, University of California, California, 1983. 50

[225] Etienne B. Roesch, Lucas Tamarit, Lionel Reveret, Didier Grandjean, David Sander, and K. R. Scherer. FACSGen: A Tool to Synthesize Emotional Facial Expressions Through Systematic Manipulation of Facial Action Units. *Journal of Nonverbal Behavior*, 35(1):1–16, November 2010. 50

[226] C. Busso and S. S. Narayanan. Joint analysis of the emotional fingerprint in the face and speech: A single subject study. *in International Workshop on Multimedia Signal Processing (MMSP*, pages 43–47, 2007. 50, 69

[227] R. Cowie, E. Douglas-Cowie, B. Apolloni, A. Romano, and W. Fellenz. What a neural net needs to know about emotion words. In *N. Mastorakis (Ed.), Computational Intelligence and Applications*, pages 109–114. World Scientific & Engineering Society Press, 1999. 51

[228] Tanja Bänziger, Véronique Tran, and K. R. Scherer. The Geneva Emotion Wheel: A tool for the verbal report of emotional reactions. *Poster presented at ISRE*, 2005. 51

[229] D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, 54(6):1063–70, June 1988. 51

[230] Humaine. D9f Usability Testing Emotion-Oriented Computing Systems : Psychometric Assessment. *Workpackage 9 Deliverable*, 2006. 51, 52

[231] S. Nowicki and M. P Duke. Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior*, 18(1):9–35, 1994. 52

[232] Albert Mehrabian. Comparison of the PAD and PANAS as models for describing emotions and for differentiating anxiety from depression. *Journal of Psychopathology and Behavioral Assessment*, 19(4):331–357, 1997. 52

[233] J. D. Morris. Observations: SAM: The Self-Assessment Manikin; An Efficient Cross-Cultural Measurement Of Emotional Response. *Journal of Advertising Research*, 35(8):63, 1995. 52

[234] D. Tsonos, K. Ikospentaki, and G. Kouroupetroglou. Towards modeling of readers' emotional state response for the automated annotation of documents. *IEEE World Congress on Computational Intelligence (WCCI 2008)*, pages 3252–3259, 2008. 52

[235] Ittipan Kanluan, M. Grimm, and Kristian Kroschel. Audio-visual emotion recognition using an emotion space concept. In *Proceedings of the 16th European Signal Processing Conference (EUSIPCO)*, Lausanna, Switzerland, 2008. 52, 97

[236] M. Grimm and K. Kroschel. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 381–385, 2005. 52, 121

[237] M M Bradley and P J Lang. Measuring Emotion: the Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994. 52

[238] J. D. Morris, K. L. Strausbaugh, and M. Nthangeni. Emotional response to advertisements (or commercials) across cultures. In *Annual Conference of the American Academy of Advertising*, Vancouvar, British Columbia, 1996. 52

[239] W.F. Blewitt and Ayesh. Modeling the emotional state of an agent through fuzzy logic with reference to the Geneva emotion wheel. In *European Simulation and Modelling (ESM\'2008) Conference*, pages 279–283, Le Harve, France, 2008. 56

[240] R. Cowie, E. Douglas-Cowie, I. Sneddon, Anton Batliner, and Catherine Pelachaud. Principles and History. In P. Petta, C. Pelachaud, and R. Cowie, editors, *Emotion-Oriented Systems: The Humaine Handbook*, chapter 2, pages 167–196. Springer Berlin Heidelberg, 2011. 57, 74, 87, 144

[241] D. Ververidis and C. Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, September 2006. 57, 87, 88, 89, 106, 107, 108, 109, 110, 113, 114, 115, 116

[242] P. Juslin and P. Laukka. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological bulletin*, 129(5):770–814, September 2003. 57, 70, 75, 87, 107, 108, 109, 110, 115

[243] P. Roach, Richard Stibbard, Jane Osborne, Simon Arnfield, and Jane Setter. Transcription of Prosodic and Paralinguistic Features of Emotional Speech. *Journal of the International Phonetic Association*, 28(1-2):83–94, 1998. 57

[244] E. Douglas-Cowie, R. Cowie, and M Schröder. A new emotion database: considerations, sources and scope. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000. 57, 65

[245] N. Campbell. A language-resources approach to emotion: corpora for the analysis of expressive speech. In *The Workshop Programme Corpora for Research on Emotion and Affect Tuesday 23 rd May 2006*, page 1, 2006. 58, 65

[246] Iris B Mauss and Michael D Robinson. Measures of emotion: A review. *Cognition & emotion*, 23(2):209–237, February 2009. 60

[247] M Schröder. Emotional speech synthesis: A review. In *Proceedings of Eurospeech*, volume 1, pages 561–564, Aalborg, Denmark, 2001. 61, 77

[248] J Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960. 61

[249] C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, 2005. 61, 82

[250] V Sacharin, K Schlegel, and K. R. Scherer. Geneva Emotion Wheel Rating study. Technical report, University of Geneva, Swiss Center for Affective Sciences, Geneva, Switzerland, 2012. 61

[251] Florian Eyben, A. Batliner, B. Schuller, Dino Seppi, and S. Steidl. Cross-Corpus classification of realistic emotionssome pilot experiments. In *Proceedings of the 3rd International Workshop on Emotion (Satellite of LREC): Corpora for Research on Emotion and Affect*, pages 77–82, 2010. 62, 121

[252] T. Eerola and J. K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, August 2010. 63

[253] I Siegert, R Bock, and B Vlasenko. Appropriate emotional labelling of non-acted speech using basic emotions, geneva emotion wheel and self assessment manikins. In *Interna-*

tional Conference on Multimedia and Expo (ICME 2011), pages 1–6, Barcalona, Spain, 2011. IEEE. 63, 97

[254] A. Batliner, S. Steidl, Christian Hacker, E. Nöth, and H. Niemann. Private emotions vs. social interaction-towards new dimensions in research on emotion. In *Proceedings of a Workshop on Adapting the Interaction Style to Affective Factors, 10th International Conference on user Modelling*, page 8, 2005. 63

[255] C. Yu, P. M. Aoki, and A. Woodruff. Detecting user engagement in everyday conversations. In *the 8th International Conference on Spoken Language Processing (ICSLP 2004)*, Jeju Island, Korea, 2004. 63

[256] R. Cowie. Perceiving emotion: towards a realistic understanding of the task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3515–3525, 2009. 63

[257] Henry S. Cheang and Marc D. Pell. The sound of sarcasm. *Speech Communication*, 50(5):366–381, May 2008. 64, 83

[258] J. Laver. *Principles of phonetics*. Cambridge University Press, Cambridge, UK, 1994. 68, 110, 111

[259] Z Zeng, M. Pantic, G I Roisman, and T. S. Huang. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009. 69

[260] T. Bänziger, D. Grandjean, and K. R. Scherer. Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT). *Emotion*, 9(5):691–704, 2009. 69, 88

[261] A. Megrabian. Communication with words. *Psychology Today*, 2(4):53–56, 1968. 69

[262] N Ambady and R Rosenthal. Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis. *Psychological bulletin*, 1992. 69

[263] Sally Planalp. Communicating emotion in everyday life: Cues, channels, and processes. In P. A. Andersen and L. K. Guerrero, editors, *Handbook of communication and emotion*, chapter 2, pages 29–48. Academic Press, San Diego, 1996. 69

[264] R. van Bezooijen. *The characteristics and recognizability of vocal expression of emotions*. Foris publications, Drodrecht, The Netherlands, 1984. 69

[265] R. W. Frick. Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97:412–429, 1985. 69

[266] K. R. Scherer. Nonlinguistic vocal indicators of emotion and psychopathology. *Emotions in personality and psychopathology*, pages 493–529, 1979. 69

[267] R. Standke. *Methoden der digitalen Sprachverarbeitung in der vokalen Kommunikationsforschung[Methods of digital speech analysis in research on vocal communication]*. Peter Lang, Frankfurt, Germany, 1993. 69

[268] K. R. Scherer, H. G. Wallbott, and A. B. Summerfield. *Experiencing emotion: A cross-cultural study*. Cambridge University Press, Cambridge, UK, 1986. 69

[269] T. Johnstone and K. R. Scherer. Vocal Communication of Emotion. In *Handbook of emotions*, volume 2, chapter 14, pages 220–235. New York: Guilford Press, 2000. 70, 88, 92, 100, 109

[270] S Wu, TH Falk, and WY Chan. Automatic Recognition of Speech Emotion using Long-term Spectro-temporal Features. In IEEE, editor, *Proceedings of the 16th International Conference on Digital Signal Processing*, Santorini-Hellas, 2009. 70

[271] E. Brunswik. *Perception and the Representative Design of Psychological Experiments*. University of California Press, Berkeley, 1956. 70

[272] K. R. Hammond and T. R. Stewert. *The Essential Brunswik: Beginnings, Explications, Applications*. Oxford University Press, New York, 2001. 70

[273] D.J. Reynolds and R. GiVord. The sounds and sights of intelligence: a lens model channel analysis. *Personality and Social Psychology Bulletin*, 27(187-200), 2001. 70

[274] I. E. Gordon. *Theories of Visual Perception*. Psychology Press, September 2004. 70, 72

[275] Tobias Brosch, Gilles Pourtois, and David Sander. The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion*, 24(3):377–401, 2010. 71

[276] J. A. Harrigan, R. Rosenthal, and K. R. Scherer. *The new handbook of methods in nonverbal behavior research*. Oxford University Press, Oxford, 2005. 75, 115

[277] H. L. Wagner. On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, 17:3–28, 1993. 75

[278] Robert Rosenthal and Donald B Rubin. Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin*, 106(2):332–337, 1989. 75

[279] Emma Rodero. Intonation and emotion: influence of pitch levels and contour type on creating emotions. *Journal of Voice: Official Journal of the Voice Foundation*, 25(1):25–34, January 2011. 76

[280] Kurt Hammerschmidt and Uwe Jürgens. Acoustical correlates of affective prosody. *Journal of voice*, 21(5):531–40, September 2007. 76, 82, 99

[281] Jackson Liscombe, Jennifer Venditti, and Julia Hirschberg. Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of Eurospeech*, pages 725–728, Geneva, Swizterland, 2003. 76, 115

[282] C. Gobl and AN Chasaide. Testing affective correlates of voice quality through analysis and resynthesis. *ISCA Tutorial and Research Workshop (ITRW)*, 2000. 77, 111

[283] SJL Mozziconacci and DJ Hermes. A study of intonation patterns in speech expressing emotion or attitude: Production and perception. *IPO Annual Progress Report*, pages 154–160, 1997. 77, 105

[284] Mirja Ilves and Veikko Surakka. Subjective responses to synthesised speech with lexical emotional content: the effect of the naturalness of the synthetic voice. *Behaviour & Information Technology*, 32(2):117–131, February 2013. 78

[285] M Ilves, V Surakka, and T Vanhala. The effects of emotionally worded synthesized speech on the ratings of emotions and voice quality. In *Affective Computing and Intelligent Interaction*, pages 588–598, 2011. 78, 84

[286] M Schröder, Anna Hunecke, and S. Krstulovic. OpenMary-Open Source Unit Selection as the Basis for Research on Expressive Synthesis. In *Blizzard Challenge Workshop*, volume 6, Pittsburgh, PA, USA, 2006. 78

[287] M Schröder. Expressing degree of activation in synthetic speech. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1128–1136, July 2006. 78, 110, 153

[288] K. R. Scherer. Expression of Emotion in Voice and Music. *Journal of Voice: Official Journal of the Voice Foundation*, 9(3):235–48, September 1995. 78

[289] Mihoko Teshigawara, N. Amir, Ofer Amir, E.M. Wlosko, and Meital Avivi. Effects of Random Splicing on Listeners' Perceptions. In *16th International Congress of Phonetic Sciences (ICPhS)*, pages 2101–2104, Saarbrücken, 2007. 79, 84, 184

[290] K. R. Scherer. Randomized splicing: A note on a simple technique for masking speech content. *Journal of Experimental Research in Personality*, 1971. 79

[291] Murray J. Munro, Tracey M. Derwing, and Clifford S. Burgess. Detection of nonnative speaker status from content-masked speech. *Speech Communication*, 52(7-8):626–637, July 2010. 79, 84

[292] F. H. Knower. Analysis of some experimental variations of simulated vocal expressions of the emotions. *The Journal of Social Psychology*, 14:369–372, 1941. 79

[293] P.L. Rogers, K. R. Scherer, and R. Rosenthal. Content filtering human speech: A simple electronic system. *Behavioral Research Methods and Instrumentation*, 3:16–18, 1971. 79, 85, 86

[294] K. R. Scherer, S. Feldstein, R.N. Bond, and R. Rosenthal. Vocal Cues to Deception: A Comparative Channel Approach. *Journal of Psycholinguistic Research*, 14(4):409–425, 1985. 79, 84

[295] a G Levitt. Reiterant speech as a test of non-native speakers' mastery of the timing of French. *The Journal of the Acoustical Society of America*, 90(6):3008–18, December 1991. 79

[296] B. Yang and M. Lugger. Emotion recognition from speech signals using new harmony features. *Signal Processing*, 90(5):1415–1423, May 2010. 82, 99, 106, 116

[297] Dmitri Bitouk, Ragini Verma, and Ani Nenkova. Class-level spectral features for emotion recognition. *Speech Communication*, 52(7-8):613–625, July 2010. 82, 99

[298] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Gerhard Rigoll, and Andreas Wendemuth. Acoustic emotion recognition: A benchmark comparison of performances. In *Workshop on Automatic Speech Recognition & Understanding*, pages 552–557. IEEE, December 2009. 82, 99

[299] Wentao Gu and Tan Lee. Quantitative analysis of F0 contours of emotional speech of Mandarin. *Proc. 6th ISCA Speech Synthesis Workshop*, pages 228–233, 2007. 82, 99

[300] Shunji Mitsuyoshi, Fuji Ren, Yasuto Tanaka, and Shingo Kuroiwa. Non-verbal Voice Emotion Analysis System. *Internation Journal of Innovative Computing, Information and Control (ICIC)*, 2(4):819–830, 2006. 82, 99

[301] Pierre-Yves Oudeyer. Novel Useful Features and Algorithms for the Recognition of Emotions in Human Speech. In *Proceedings of the 1st International Conference on Speech Prosody*, pages 3–6, Aix-en-Provence, France, 2002. 82, 99

[302] Jiahong Yuan, Liqin Shen, and Fangxin Chen. The acoustic realization of anger, fear, joy and sadness in Chinese. In *Proceedings of ICSLP*, pages 2025–2028, Denver, Colorado, USA, 2002. 82, 99

[303] S Mcgilloway, R. Cowie, E. Douglas-Cowie, S Gielen, M Westerdijk, and S Stroeve. Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark. *ISCA Workshop on Speech and Emotion, Belfast*, 2000. 82, 99, 104, 105

[304] SJL Mozziconacci and DJ Hermes. Role of intonation patterns in conveying emotion in speech. In *Proceedings of International Congress of Phonetic Sciences (ICPhS)*, pages 2001–2004, San Francisco, USA, 1999. 82, 99

[305] E Zetterholm. Emotional speech focusing on voice quality. *Fonetik 99: the Swedish Phonetics Conference (Gothenburg papers in theoretical linguistics)*, 81:145–148, 1999. 82, 99

[306] A Paeschke, M Kienast, and WF Sendlmeier. F0-Contours in Emotional Speech. In *Proceedings of the 14th International Conference of Phonetic Sciences*, volume 2, pages 929–932, San Francisco, USA, 1999. 82, 99

[307] Gözde Özbal and Daniele Pighin. Evaluating the impact of syntax and semantics on emotion recognition from text. *Proceedings of the 14th international conference on Computational Linguistics and Intelligent Text Processing (CICLing'13)*, 2(161-173), 2013. 82

[308] Edward Chao-Chun Kao, Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh, and Von-Wun Soo. Towards Text-based Emotion Detection A Survey and Possible Improvements. In *International Conference on Information Management and Engineering (ICIME)*, pages 70–74. Ieee, 2009. 82

[309] Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. Emotion recognition from text using semantic labels and separable mixture models. *ACM Transactions on Asian Language Information Processing*, 5(2):165–183, June 2006. 82

[310] Richard Craggs and Mary Mcgee Wood. A categorical annotation scheme for emotion in the linguistic content of dialogue. *Affective Dialogue Systems*, 3068:89–100, 2004. 82, 151, 152

[311] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, pages 577–580, 2004. 82

[312] Ronald Müller, Björn Schuller, and Gerhard Rigoll. Enhanced robustness in speech emotion recognition combining acoustic and semantic analyses. In *Workshop "From Signals To Signs of Emotion and Vice Versa", EU-IST FP6 Network of Excellence HUMAINE*, Santorini, Greece, 2004. 82

[313] A. Batliner. How to find trouble in communication. *Speech Communication*, 40(1-2):117–143, April 2003. 82, 95

[314] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of ICSLP*, number June, pages 2037–2040, 2002. 82

[315] CM Lee, SS Narayanan, and Roberto Pieraccini. Combining acoustic and language information for emotion recognition. *Neural Computation*, 2002:6–9, 2002. 82

[316] Jackson Liscombe and Giuseppe Riccardi. Using context to improve emotion detection in spoken dialog systems. In *Proceedings of Interspeech'05*, Lisbon, Portugal, 2005. 82

[317] J. Bertels, R Kolinsky, and J. Morais. Norms of Emotional Valence, Arousal, Threat Value and Shock Value for 80 Spoken French Words: Comparison Between Neutral and Emotional Tones of Voice. *Psychologica belgica*, 49(1):19–40, 2009. 83, 84

[318] Juliane Degner. Affective priming with auditory speech stimuli. *Language and Cognitive Processes*, 26(10):1710–1735, December 2011. 84

[319] Marc D Pell, Abhishek Jaywant, Laura Monetta, and Sonja a Kotz. Emotional speech processing: disentangling the effects of prosody and semantic cues. *Cognition & emotion*, 25(5):834–53, August 2011. 84

[320] Lynne C Nygaard and Jennifer S Queen. Communicating emotion: linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception and Performance*, 34(4):1017–30, August 2008. 84, 184

[321] Keiko Ishii, Jose Alberto Reyes, and Shinobu Kitayama. Spontaneous attention to word content versus emotional tone: differences among three cultures. *Psychological Science*, 14(1):39–46, January 2003. 84, 128, 167

[322] S.G. Koolagudi, Ramu Reddy, Jainath Yadav, and K.S. Rao. IITKGP-SEHSC: Hindi speech corpus for emotion analysis. In *International Conference on Devices and Communications (ICDeCom)*, pages 1–5. IEEE, 2011. 84

[323] J. Kim. Bimodal emotion recognition using speech and physiological changes. *Robust Speech Recognition and Understanding*, pages 265–280, 2007. 84

[324] Kala Lakshminarayanan, Dorit Ben Shalom, Virginie van Wassenhove, Diana Orbelo, John Houde, and David Poeppel. The effect of spectral manipulations on the identification of affective and linguistic prosody. *Brain and Language*, 84(2):250–263, February 2003. 84

[325] R. J. McNally, M. W. Otto, and C. D. Hornig. The voice of emotional memory: content-filtered speech in panic disorder, social phobia, and major depressive disorder. *Behaviour research and therapy*, 39(11):1329–37, November 2001. 84, 85, 86

[326] K. R. Scherer, J. Koivumaki, and R. Rosenthal. Minimal Cues in the Vocal Communication of Affect: Judging Emotions from Content-Masked Speech. *Journal of Psycholinguistic*, 1(3):269–285, 1972. 85, 88, 182

[327] Deborah a. Vickers, Brian C. J. Moore, and Thomas Baer. Effects of low-pass filtering on the intelligibility of speech in quiet for people with and without dead regions at high frequencies. *The Journal of the Acoustical Society of America*, 110(2):1164, 2001. 86

[328] M. Ardoint and C. Lorenzi. Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues. *Hearing research*, 260(1-2):89–95, February 2010. 86

[329] Gaetan Gilbert and Christian Lorenzi. The ability of listeners to use recovered envelope cues from speech fine structure. *The Journal of the Acoustical Society of America*, 119(4):2438, 2006. 86

[330] Stanford W Gregory, Will Kalkhoff, Sarah K Harkness, and Jessica L Paull. Targeted high and low speech frequency bands to right and left ears respectively improve task performance and perceived sociability in dyadic conversations. *Laterality*, 14(4):423–40, July 2009. 86, 168

[331] Julia K MacCallum, Aleksandra E Olszewski, Yu Zhang, and Jack J Jiang. Effects of low-pass filtering on acoustic analysis of voice. *Journal of Voice: Official Journal of the Voice Foundation*, 25(1):15–20, January 2011. 86, 94

[332] B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 Emotion Challenge. In *10th Annual Conference of the International Speech Communication Association*, pages 312–315, Brighton, United Kingdom, 2009. 86, 95

[333] N. Amir, O. Kerret, and D. Karlinski. Classifying emotions in speech: a comparison of methods. In *Proceedings of European conference on speech communication and technology (EUROSPEECH'01)*, Aalborg, Denmark, 2001. 86

[334] RW Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence - Graph Algorithms and Computer Vision*, 23(10):1174–1191, 2001. 86, 89

[335] L. Devillers, J.C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner. Workshop on Corpora for Research on Emotion and Affect. In *6th International Conference on Language Resources and Evaluation (LREC'2008)*, Marrackech, Morrocco, 2008. 87

[336] Julia Hirschberg, Catherine Pelachaud, David Sadek, Hannes Pirker, K. R. Scherer, L. Devillers, Samuel Ron, Stephania Balzarotti, F Manzoni, and Olivier Rosec. The Workshop Programme Corpora for Research on Emotion and Affect. In *3rd International Workshop on EMOTION (satellite of LREC)*, 2006. 87

[337] Ernest Kramer. Elimination of verbal cues in judgments of emotion from voice. *Journal of Abnormal and Social Psychology*, 68(4):390 – 396, 1964. 88

[338] Thurid Vogt and E André. Comparing Feature Sets for Acted and Spontaneous Speech in view of Automatic Emotion Recognition. In *International Conference on Multimedia and Expo (ICME)*, pages 474–477, Amsterdam, 2005. IEEE. 88, 95

[339] C. Cullen and B. Vaughan. Emotional Speech Corpora for Analysis and Media Production. In *3rd International Conference on Semantic and Digital Media Technologies (SAMT)*, Koblenz, Germany, 2008. 89, 96

[340] T. Johnstone. Emotional speech elicited using computer games. In *Fourth International Conference on Spoken Language Processing (ICSLP'96)*, volume 3, pages 1985 – 1988, Philadelphia, PA, 1996. 89, 90

[341] C Johns-Lewis. Prosodic differentiation of discourse modes. *Intonation in discourse*, pages 199–220, 1986. 89

[342] C. Busso and S. S. Narayanan. Scripted dialogs versus improvisation: lessons learned about emotional elicitation techniques from the IEMOCAP database. In *Interspeech*, pages 1670–1673, Brisbane, Australia, 2008. 89

[343] B. Schuller, A. Batliner, Dino Seppi, S. Steidl, T. Vogt, Johannes Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and Others. The Relevance of Feature Type for the Automatic Classification of Emotional User States : Low Level Descriptors and Functionals. In *Proceedings of INTERSPEECH 2007*, volume 101, pages 2253–2256, 2007. 89, 95

[344] A. Gerrards-Hesse, K. Spies, and F.W. Hesse. Experimental inductions of emotional states and their effectiveness: A review. *British Journal of Psychology*, 85(1):55–78, 1994. 91

[345] A. S. Göritz. The Induction of Mood via the WWW. *Motivation and Emotion*, 31(1):35–47, November 2006. 91

[346] F. Weiss, G. S. Blum, and L. Gleberman. Anatomically based measurement of facial expressions in simulated versus hypnotically induced affect. *Motivation and Emotion*, 11(1):67–81, March 1987. 91

[347] E. Velten. A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 6(4):473–482, 1968. 91

[348] J. Rottenberg, R. D. Ray, and J. J. Gross. Emotion elicitation using films. *The handbook of emotion elicitation and assessment*, pages 9–28, 2007. 91

[349] S. K. De l'Etoile. The effect of a musical mood induction procedure on mood state-dependent word retrieval. *Journal of music therapy*, 39(2):145, 2002. 91

[350] B Ruffle. Gift giving with emotions. *Journal of Economic Behavior & Organization*, 39(4):399–420, July 1999. 91

[351] L. Nummenmaa and P. Niemi. Inducing affective states with success–failure manipulations: A meta-analysis. *Emotion*, 4(2):207 – 214, 2004. 92

[352] S. Steidl. *Automatic classification of emotion-related user states in spontaneous childrens speech*. PhD thesis, Universität Erlangen-Nürnberg, 2009. 92

[353] H. Helfrich, R. Standke, and K. R. Scherer. Vocal indicators of psychoactive drug effects. *Speech Communication*, 3(3):245 – 252, 1984. 92

[354] T. Johnstone, C. M. van Reekum, K. Hird, K. Kirsner, and K. R. Scherer. Affective Speech Elicited With a Computer Game. *Emotion*, 5(4):513–518, 2005. 92

[355] R. Kehrein. The prosody of authentic emotions. In *Proceedings of Speech Prosody*, pages 423–426, Aix-en-Provence, France, 2002. 92

[356] Jennifer Barrett and Tomáš Paus. Affect-induced changes in speech production. *Experimental Brain Research*, 146(4):531–537, 2002. 92

[357] Tom Johnstone, Carien M van Reekum, Kathryn Hird, Kim Kirsner, and K. R. Scherer. Affective speech elicited with a computer game. *Emotion (Washington, D.C.)*, 5(4):513–8, December 2005. 92

[358] R Westermann, K Spies, G Stahl, and F. W. Hesse. Relative effectiveness and validity of mood induction procedures: a meta-analysis. *European Journal of Social Psychology*, 26:557–580, 1996. 92

[359] U. Jürgens. Neural pathways underlying vocal control. *Neuroscience and Biobehavioral Reviews*, 26:235–58, 2002. 93

[360] P. Ekman. Biological and cultural contributions to body and facial movement. In John Blacking, editor, *The anthropology of the body*, pages 39–84. Academic Press, London, 1977. 93

[361] J. R. Krebs and R. Dawkins. Animal signals: mind-reading and manipulation. In J.R. Krebs and N.B. Davies, editors, *Behavioural ecology: an evolutionary approach*, pages 380–40. Oxford: Blackwell, 2nd edition, 1984. 93

[362] K. R. Scherer. Vocal Measurement of Emotion. *Emotion: Theory, research, and experience*, 1989. 93

[363] N. Campbell. Databases of Emotional speech. *ISCA Tutorial and Research Workshop on Speech and Emotion*, pages 114–121, 2000. 93

[364] C. Cullen, B. Vaughan, S. Kousidis, Y. Wang, C. McDonnell, and D. Campbell. Generation of high quality audio natural emotional speech corpus using task based mood induction. In *International Conference on Multidisciplinary Information Sciences and Technologies Extremadura (InSciT)*, Merida, Spain, 2006. 94, 96, 193

[365] C Carson, D Ingrisano, and K Eggleston. The Effect of Noise on Computer-Aided Measures of Voice: A Comparison of CSpeechSP and the Multi-Dimensional Voice Program Software Using the CSL 4300B Module and Multi-Speech for Windows. *Journal of Voice*, 17(1):12–20, March 2003. 94

[366] Dimitar D Deliyski, Heather S Shaw, and Maegan K Evans. Adverse effects of environmental noise on acoustic voice quality measurements. *Journal of voice*, 19(1):15–28, March 2005. 94

[367] V Parsa, D G Jamieson, and B R Pretty. Effects of Microphone Type on Acoustic Measures of Voice. *Journal of Voice: Official Journal of the Voice Foundation*, 15(3):331–43, September 2001. 94

[368] S Drgas and M. A. Blaszak. Perception of speech in reverberant conditions using AM-FM cochlear implant simulation. *Hearing research*, 269(1-2):162–8, October 2010. 94

[369] R. Cowie, E. Douglas-Cowie, M. McRorie, I. Sneddon, L. Devillers, and N. Amir. Issues in Data Collection. In P. Petta, C. Pelachaud, and R. Cowie, editors, *Emotion-Oriented Systems: The Humaine Handbook*, pages 197–213. Springer Berlin Heidelberg, 2011. 95

[370] Janneke Wilting, Emiel Krahmer, and Marc Swerts. Real vs. acted emotional speech. In *Proceedings of Interspeech*, pages 805–808, 2006. 95

[371] I Murray and J Arnott. Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. *Computer Speech & Language*, 22(2):107–129, April 2008. 95

[372] D Ververidis and C Kotropoulos. Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections. *Proceedings of European Signal Processing Conference (EUSIPCO 2006)*, 2006. 95

[373] B. Vaughan, S. Kosidis, C. Cullen, and Y. Wang. Task-based mood induction procedures for the elicitation of natural emotional responses. In *the 4th International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA 2007*, Orlando, Florida., 2007. 96, 193

[374] E. Douglas-Cowie. Data and Databases. In P. Petta, C. Pelachaud, and R. Cowie, editors, *Emotion-Oriented Systems: The Humaine Handbook*, pages 163–212. Springer Berlin Heidelberg, 2011. 96

[375] Jun-Heng Yeh, Tsang-Long Pao, Ching-Yi Lin, Yao-Wei Tsai, and Yu-Te Chen. Segment-based emotion recognition from continuous Mandarin Chinese speech. *Computers in Human Behavior*, 27(5):1545–1552, September 2011. 97, 117

[376] M. Grimm, Kristian Kroschel, and S. S. Narayanan. Support vector regression for automatic recognition of spontaneous emotions in speech. In *IEEE International Conference*

on Acoustics, Speech and Signal Processing, 2007 (ICASSP'2007), volume 4, pages IV–1085–IV–1088. Ieee, April 2007. 97

[377] Cécile Pereira. Dimensions of emotional meaning in speech. ... *Research Workshop (ITRW) on Speech and Emotion*, pages 1–4, 2000. 97, 110

[378] M Tatham and K Morton. *Expression in speech: Analysis and Synthesis*. Oxford University Press, Oxford, UK, 2004. 97, 157

[379] K. R. Scherer, H. Wagner, and A. Manstead. Vocal correlates of emotional arousal and affective disturbance. In H. Wagner and A. Manstead, editors, *Handbook of Psychophysiology: Emotion and social behavior*, chapter 7, pages 165–197. Wiley, London, UK, 1989. 99, 100

[380] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, 1970. 100

[381] Mikael Nilsson and Marcus Ejnarsson. Speech Recognition using Hidden Markov Model performance evaluation in noisy environment. *Signal Processing*, 2002. 101

[382] Harry Hollien and G Paul Moore. Measurements of the Vocal Folds during Changes in Pitch. *Journal of Speech, Language, and Hearing Research*, 3(2):157–165, 1960. 101

[383] C. Gobl and Ailbhe Ni Chasaida. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1-2):189–212, 2003. 102, 111, 112

[384] Stefan Werner and Eric Keller. Prosodic aspects of speech. In *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges*, chapter 2, pages 23–40. John Wiley, Chichester, 1994. 103, 104, 106, 107, 196

[385] T. Dutoit. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997. 103

[386] J. Holmes and W. Holmes. *Speech Synthesis and Recognition*. Taylor & Francis, 11 New Fetter Lane, London, 2 edition, 2001. 103, 114

[387] Anne Cutler, Delphine Dahan, and Wilma Van Donselaar. Prosody in the comprehension of spoken language: A literature review. *Language & Speech*, 40(2):141, 1997. 103

[388] Jianhua Tao, Yongguo Kang, and Aijun Li. Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1145–1154, 2006. 104

[389] I. R. Murray and J. L. Arnott. Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93:1097–1108, 1993. 104, 111, 112

[390] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. TOBI: A standard for labeling english prosody. *2nd International Conference on Spoken Lanuage Processing (ICSLP 92)*, pages 867–870, 1992. 104

[391] Je Hun Jeon and Yang Liu. Semi-supervised learning for automatic prosodic event detection using co-training algorithm. *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 540–548, 2009. 105

[392] C. Cullen, B. Vaughan, and S. Kousidis. LinguaTag: an emotional speech analysis application. *12th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI '08)*, 2008. 105

[393] J. 't Hart, R. Collier, and A. Cohen. *A perceptual study of intonation: An experimental phonetic approach to speech melody*. Cambridge University Press, 1990. 105

[394] Paul Taylor. The rise/fall/connection model of intonation. *Speech Communication*, 15(1-2):169–186, October 1994. 105

[395] Robert A J Clark. *Generating Synthetic Pitch Contours Using Prosodic Structure*. PhD thesis, University of Edinburgh, 2003. 105

[396] A. Cruttenden. *Intonation*. Cambridge University Press, Cambridge, UK, 1997. 106

[397] R. Cowie and M Schröder. Piecing together the emotion jigsaw. *Machine Learning for Multimodal Interaction*, pages 305–317, 2005. 106

[398] Sam Tilsen and Keith Johnson. Low-frequency Fourier analysis of speech rhythm. *The Journal of the Acoustical Society of America*, 124(2):34–39, 2008. 107

[399] Iker Luengo, Eva Navas, and Inmaculada Hernáez. Combining spectral and prosodic information for emotion recognition in the interspeech 2009 emotion challenge. In *10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, pages 332–335, Brighton, United Kingdom, 2009. 108

[400] Jürgen Trouvain and WJ Barry. The prosody of excitement in horse race commentaries. *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 86–91, 2000. 108

[401] F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 3:1970–1973, 1996. 108

[402] Akemi Iida, Nick Campbell, and Soichiro Iga. A Speech Synthesis System with Emotion for Assisting Communication. In *ITRW on Speech and Emotion*, Newcastle, Northern Ireland, 2000. 108

[403] B. Heuft, T. Portele, and M. Rauth. Emotions in time domain synthesis. In *Proc. Int. Conf. Spoken Language Processing (ICSLP '96)*, volume 3, pages 1974–1977, 1996. 108

[404] Martti Vainio and Toomas Altosaar. Modeling the microprosody of pitch and loudness for speech synthesis with neural networks. In *Proceedings of the 5th International Conference on Spoken Language Processing*, pages 1–4, Sydney, Australia, 2000. Citeseer. 110

[405] J. Laver. *The phonetic description of voice quality*. Cambridge University Press, Cambridge, England, 1980. 110, 111, 114

[406] E. Zetterholm. Prosody and voice quality in the expression of emotions. In *Proceedings of the Seventh Australian International Conference on Speech Science and Technology (SST)*, pages 109–113, Sydney, Australia, 1998. 110, 111

[407] Sonja Biersack and Vera Kempe. Tracing vocal emotion expression through the speech chain: Do listeners perceive what speakers feel? In *ISCA Workshop on Plasticity in Speech Perception*, pages 211–214, London, UK, 2005. 110

[408] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic*, 17:97–110, 1993. 111

[409] CT Ferrand. Harmonics-to-Noise Ratio: An Index of Vocal Aging. *Journal of Voice*, 16(4):480–487, 2002. 111

[410] A. Ozdas, R.G. Shiavi, S.E. Silverman, M.K. Silverman, and D.M. Wilkes. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Biomedical Engineering*, 51(9):1530–1540, 2004. 111

[411] C. Gobl, E. Bennet, and N. Chasaide. Expressive synthesis: how crucial is voice quality. In *Proceedings of IEEE Workshop on Speech Synthesis*, pages 91–94, 2002. 111

[412] G. Klasmeyer and W. Sendlmeier. Objective voice parameters to characterise the emotional content in speech. In *Proceedings of the 13th International Congress of Phonetic Sciences*, volume 2, pages 182–185, Stockholm, Sweden, 1995. 111

[413] A.-M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen. Physical variation related to stress and emotionally state: a preliminary study. *Journal of Phonetics*, 24:313–335, 1996. 111, 112

[414] F. Burkhardt and W.F. Sendlmeier. Verification of acoustical correlates of emotional speech using formant-synthesis. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Citeseer, 2000. 111

[415] J. Krajewski and B. Kröger. Using prosodic and spectral characteristics for sleepiness detection. In *8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, volume 8, pages 1841–1844, Antwerp, Belgium, 2007. 113, 115

[416] G. Zhou, J.H.L. Hansen, and J.F. Kaiser. Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 9(3):201–216, March 2001. 113, 116

[417] Melissa Ann Epstein. *Voice Quality and Prosody*. PhD thesis, University of California, Los Angeles, 2002. 114, 115

[418] D. Datcu and L. J. M. Rothkrantz. The recognition of emotions from speech using gentleboost classifier. a comparison approach. *International Conference on Computer Systems and Technologies*, 2006. 115

[419] Muharram Mansoorizadeh and Nasrollah.M. Charkari. Speech Emotion Recognition: Comparison of Speech Segmentation Approaches. In *Proceedings of IKT'07*, Mashad, Iran, 2007. 115

[420] D. Ververidis and C. Kotropoulos. Emotional speech classification using gaussian mixture models. *IEEE International Symposium on Circuits and Systems (ISCAS)*, 3:2871 – 2874, 2005. 115

[421] D. Ververidis, C. Kotropoulos, and I. Pitas. Automatic emotional speech classification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 593–596, Montreal, Canada, 2004. 115

[422] L. Vidrascu and L. Devillers. Detection of real-life emotions in call centers. *Proceedings of Interspeech*, pages 1841–1844, 2005. 115

[423] Bogdan Vlasenko, B. Schuller, Andreas Wendemuth, and G. Combining frame and turn-level information for robust recognition of emotions within speech. *Proceedings of Interspeech*, 2007. 115

[424] Monique Biemans. *Gender variation in voice quality*. PhD thesis, LOT: Netherlands Graduate School of Linguistics, 2000. 115

[425] Mattias Heldner. Spectral emphasis as an additional source of information in accent detection. *ISCA Tutorial and Research Workshop (ITRW)*, 2001. 115

[426] I Linnankoski, L Leinonen, M Vihla, M Laakso, and S Carlson. Conveyance of emotional connotations by a single word in English. *Speech Communication*, 45(1):27–39, January 2005. 115

[427] S Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics Speech and Signal Processing*, 28(4):357–366, 1980. 116

[428] C. M. Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, C. Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition based on phoneme classes. *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2004)*, pages 889—-892, 2004. 116

[429] C. Busso, Sungbok Lee, and S. Narayanan. Using neutral speech models for emotional speech analysis. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 07)*, pages 2304–2307, Antwerp, Belgium, 2007. 116

[430] W. Hu, H. and Xu, M.-X. and Wu. GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition. In *Acoustics, Speech and Signal Processing (ICASSP'2007)*, 2007. 116

[431] V. Sethu, E. Ambikairajah, and J. Epps. Group delay features for emotion detection. In *Proceedings of INTERSPEECH 2007*, pages 2273–2276, Antwerp, Belgium, 2007. 116

[432] T Nwe. Speech emotion recognition using hidden Markov models. *Speech Communication*, 41(4):603–623, November 2003. 116

[433] M. Lugger and B. Yang. *Psychological motivated multi-stage emotion classification exploiting voice quality features*, chapter 22. In-Tech, 2008. 116

[434] D. Ververidis and C Kotropoulos. Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. *Signal Processing*, 88(12):2956–2970, December 2008. 118

[435] Jia Rong, Gang Li, and Yi-Ping Phoebe Chen. Acoustic feature selection for automatic emotion recognition from speech. *Information Processing & Management*, 45(3):315–328, 2009. 118

[436] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, November 2008. 121, 137, 152

[437] S.G. Hart and L.E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload*, 1:139–183, 1988. 127

[438] Daniel Grühn and Jacqui Smith. Characteristics for 200 words rated by young and older adults: age-dependent evaluations of German adjectives (AGE). *Behavior research methods*, 40(4):1088–97, November 2008. 128, 167

[439] Mireille Besson and Cyrille Magne. Emotional prosody: sex differences in sensitivity to speech melody. *Trends in Cognitive Sciences*, 6(10):405–407, 2002. 128, 167, 168

[440] M. Nilsenová, Martijn Goudbeek, and Luuk Kempen. The Relation Between Pitch Perception Preference and Emotion Identification. *Interspeech*, pages 110–113, 2010. 128, 167, 168

[441] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Charles Florin, Luca Bogoni, and Linda Moy. Learning From Crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010. 129

[442] C. Cullen, B. Vaughan, S. Kousidis, and J. McAuley. Emotional Speech Corpus Construction, Annotation and Distribution. In *the sixth international conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008. 133, 134

[443] J. Snel, A. Tarasov, C. Cullen, and SJ Delany. A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpus. In *4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES3 2012)*, Istanbul, Turkey, 2012. 140

[444] M. Grimm, K. Kroschel, and S. Narayanan. The Vera am Mittag German Audio-Visual Emotional Speech Database. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 865–868, Hannover, Germany, 2008. 150

[445] K.P. Truong, M.A. Neerincx, and D.A. van Leeuwen. Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data. In *9th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 318–321, Brisbane, Australia, 2008. 151

[446] Dennis Reidsma and DKJ Heylen. Annotating Emotion in Meetings. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1117–1122, Genoa, Italy, 2006. 151

[447] K Koffka. *Principles of Gestalt Psychology*. Psychology Press, 1999. 157

[448] C De Looze and DJ Hirst. Detecting changes in key and range for the automatic modelling and coding of intonation. *Speech Prosody*, 2008. 166

[449] Paul Boersma and David Weenink. Praat: doing phonetics by computer, 2012. 166, 195

[450] TakashiX. Fujisawa and Kazuyuki Shinohara. Sex differences in the recognition of emotional prosody in late childhood and adolescence. *The Journal of Physiological Sciences*, 61(5):429–435, 2011. 168

[451] A Schirmer and SA Kotz. Sex Differentiates the Stroop-effect in Emotional Speech: ERP Evidence. In *Proceedings Speech Prosody*, pages 631–634, Aix-en-Provence, France, 2002. 168

[452] IBM Corp. IBM SPSS Statistics (Version 20), 2011. 168

[453] Sarah Boslaugh and Paul A Watters. *Statistics in a Nutshell: A Desktop Quick Reference*. O'Reilly Media, July 2008. 176

[454] K. R. Scherer. 4. Methods of research on vocal communication: paradigms and parameters. *affective-sciences.org*, pages 137–198, 1982. 195

[455] A. L. Gilet. Mood induction procedures: a critical review. *L'Encéphale*, 34(3):233–9, July 2008. 198

Appendices

A

Preliminary Surveys

Questionnaire

Number of participants: **7**

Procedure for participants:

1. Read instructions.
2. Attend to Evaluation/Activation questionnaire.
3. Rate clips.
4. Attend to subjective workload questionnaire.

Instructions:

Q1 Which of the following is best described by Evaluation (pick one answer):

(a) A speech segment relating to measurement.

(b) A speech segment relating to examination.

(c) A speech segment that sounds like a whisper.

(d) A speech segment where the speakers voice conveys the benefit of (or problem with) something.

Q2 Which of the following is best described by Activation (pick one answer):

(a) A speech segment relating to work levels.

(b) A speech segment relating to politics.

(c) A speech segment that contains physical arousal in the voice due to emotion.

(d) A speech segment where the speaker begins an action.

Subjective workload of tool usage:

Q3 How hurried or rushed was the pace of the task?

(a) Very low (b) Low (c) Normal (d) High (e) Very high

Q4 How mentally demanding was the task?

(a) Very low (b) Low (c) Normal (d) High (e) Very high

Q5 How uncertain, irritated, and stressed were you?

(a) Very low (b) Low (c) Normal (d) High (e) Very high

Q6 If you were asked to rate a certain number of clips on a daily basis, do you think 3-7 clips should be:

(a) Enough (b) Increased (c) Decreased

Questionnaire results:

	Correct	Incorrect			
Q1	6	1			
Q2	6	1			
	Very low	Low	Normal	High	Very high
Q3	1	1	3*	2	0
Q4	0	2	4*	0	0
Q5	3*	1	2	1	0
	Kept the same	Increase	Decreased		
Q6	4*	3	0		

*Mode

Participant feedback:

A. Tool functionality:

Participant 1 (Male, technical):

- Participant found restriction on using characters within password frustrating (e.g. hyphen).
- If all fields are filled in when a new user account is created, button should be highlighted.
- Firefox 5 showed some issues with volume button in player.

Participant 3 (Female, non-technical):

- Easy to sign in, satisfied with no need for email confirmation.

Participant 4 (Male, non-technical):

- Overall easy to understand.

Participant 6 (Female, non-technical):

- Easy login, satisfied with no email confirmation.

B. Instructions:

Participant 1 (Male, technical):

- Suggested answers were more obvious because they were longer (more text). Asked how participant knew for sure it was the right answer—response was from instructions.
- Participant didnt agree that Activity and Evaluation should be two different scales. Suggested Evaluation to be binary, and a scale for Activity. Participant also questioned surprise.

Participant 2 (Male, technical):

- The word "overlap" used to describe Activity and Evaluation was confusing. Participant thought he needed to rate on a matrix.

Participant 3 (Female, non-technical):

- Instructions were clear.

Participant 6 (Female, non-technical):

- Easy to understand.

Participant 7 (Male, somewhat technical)

- Found instructions a bit overwhelming, suggested Activation should be kept shorter—found the reference to adrenalin a bit confusing with the example given i.e. receiving a gift but said there is no reference to fight or flight. Participant understood it better once the rating tool was presented (labels etc). Examples did help to understand the instructions.

C. Presented speech clips:

Participant 1 (Male, technical):

- Listener felt clips didnt portray 'real' emotion. Suggested a scale for authenticity/genuiness.
- Suggested a baseline clip to compare against.
- Suggested 7 clips to rate.

Participant 2 (Male, technical):

- Last listened to clip was too short.
- Listened to clips several times to try hear tone of voice, and not semantic content.

Participant 3 (Femal, non-technical):

- 1st speech clip to short.

Participant 5 (Female, non-technical):

- Had to listen to clips 2-3 times.
- Listened for linguistic content.

Participant 6 (Female, non-technical):

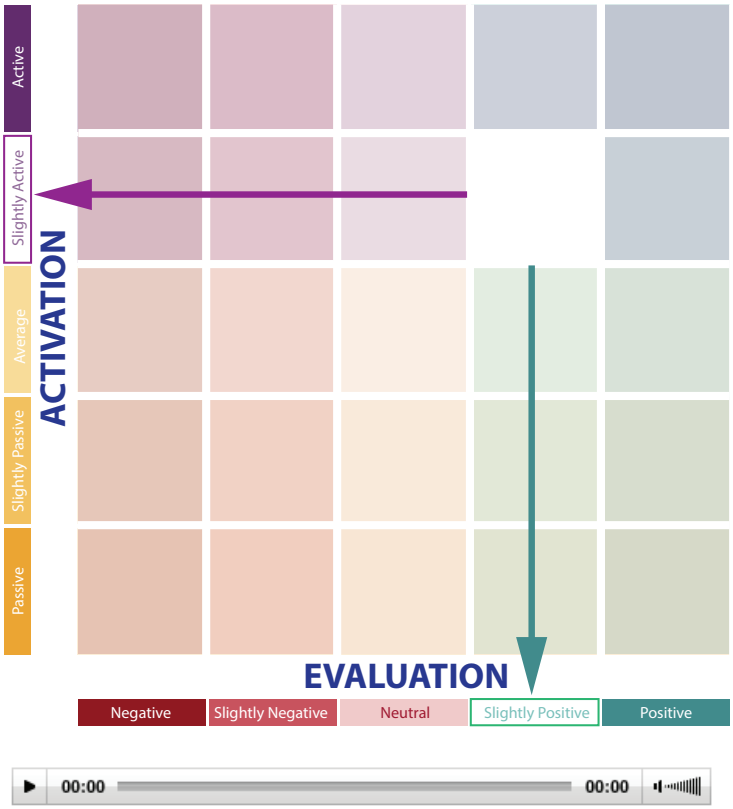
- Found the clips used "Weird".

B

Instructions: Base Line Speech Clip

Listen to Example:

In this following clip, there has been some agreement between listeners on the activation and evaluation scales. This clip has been rated as "Slightly Active" on the **Activation** scale and "Slightly Positive" on the **Evaluation** scale. It is important to note, however, that this is only given as an example and that there is no correct answer when rating the clips.



Please note:

For the listening task, you will be presented with two individual scales rather than the grid shown here to the left. The grid is used to show that both scales can overlap.

You can click on continue if you feel you understand both evaluation and activation. If you don't ask the researcher to explain.

Continue ▶

Figure B.1: Base line speech clip as part of the instructions with associated Ground Truth value.

Segmentation

SAMPLE RATE: 44100
BIT DEPTH: 24-bit

T R A C K L I S T I N G — Total: 20

MIP Session: Camilla_Christine

TRACK NAME:	Camilla		
FILE NAME	START TIME	END TIME	DURATION
B2_CC_200309	0:03.381	0:06.962	0:03.580
B3_CC_200309	0:07.004	0:08.911	0:01.906
B5_CC_200309	0:13.050	0:15.595	0:02.545
B6_CC_200309	0:18.908	0:20.726	0:01.818
B7_CC_200309	0:21.171	0:23.716	0:02.545
B8_CC_200309	0:23.837	0:26.060	0:02.222
B10_CC_200309	0:36.403	0:38.908	0:02.505
A9_CC_200309	9:03.182	9:06.415	0:03.232
A5_CC_200309	9:26.818	9:29.121	0:02.303
A4_CC_200309	9:35.000	9:36.970	0:01.970
A3_CC_200309	9:36.975	9:38.160	0:01.185
A2_CC_200309	9:38.162	9:40.467	0:02.305

TRACK NAME:	Christine		
FILE NAME	START TIME	END TIME	DURATION
B1_CC_200309	0:00.989	0:04.020	0:03.030
B4_CC_200309	0:09.333	0:12.161	0:02.828
B9_CC_200309	0:29.090	0:31.878	0:02.787
A10_CC_200309	9:02.496	9:04.980	0:02.484
A8_CC_200309	9:06.192	9:08.556	0:02.363
A7_CC_200309	9:09.000	9:13.223	0:04.222
A6_CC_200309	9:16.354	9:19.364	0:03.010
A1_CC_200309	9:38.293	9:42.475	0:04.181

Note: File start with B= Before; Filestart with A= After

T R A C K L I S T I N G — Total: 20

MIP Session: David_Luc

TRACK NAME:	David		
FILE NAME	START TIME	END TIME	DURATION
B1_DL_230309	0:00.761	0:04.896	0:04.135
B2_DL_230309	0:07.960	0:10.203	0:02.243
B5_DL_230309	0:14.459	0:15.520	0:01.061
B7_DL_230309	0:21.098	0:23.862	0:02.763
B10_DL_230309	0:27.987	0:30.066	0:02.078
A10_DL_230309	10:27.900	10:29.505	0:01.605
A7_DL_230309	10:41.275	10:43.415	0:02.140
A5_DL_230309	10:59.465	11:01.265	0:01.799
A3_DL_230309	11:04.426	11:06.517	0:02.091
A1_DL_230309	11:11.235	11:15.710	0:04.474

TRACK NAME:	Luc		
FILE NAME	START TIME	END TIME	DURATION
B3_DL_230309	0:10.143	0:11.885	0:01.742
B4_DL_230309	0:12.246	0:13.528	0:01.281
B6_DL_230309	0:16.952	0:17.894	0:00.941
B8_DL_230309	0:23.381	0:24.593	0:01.211
B9_DL_230309	0:25.944	0:28.097	0:02.152
A9_DL_230309	10:32.666	10:35.390	0:02.723
A8_DL_230309	10:38.503	10:41.469	0:02.966
A6_DL_230309	10:44.825	10:46.625	0:01.799
A4_DL_230309	11:00.632	11:02.286	0:01.653
A2_DL_230309	11:05.885	11:08.268	0:02.383

T R A C K L I S T I N G Total: 20

MIP Sesssion: Emma_Jason

TRACK NAME:	Emma		
FILE NAME	START TIME	END TIME	DURATION
B2_EJ_250509	0:10.092	0:11.289	0:01.196
B3_EJ_250509	0:11.296	0:13.854	0:02.557
B5_EJ_250509	0:15.512	0:17.067	0:01.555
B8_EJ_250509	0:25.735	0:28.177	0:02.441
B10_EJ_250509	0:35.755	0:42.449	0:06.693
A8_EJ_250509	2:55.357	2:58.008	0:02.650
A6_EJ_250509	3:02.823	3:04.437	0:01.613
A5_EJ_250509	3:04.696	3:06.790	0:02.093
A3_EJ_250509	3:08.732	3:10.537	0:01.805
A1_EJ_250509	3:13.654	3:19.589	0:05.935

TRACK NAME:	Jason		
FILE NAME	START TIME	END TIME	DURATION
B1_EJ_250509	0:05.892	0:09.156	0:03.263
B4_EJ_250509	0:14.014	0:15.713	0:01.698
B6_EJ_250509	0:17.084	0:19.253	0:02.168
B7_EJ_250509	0:21.463	0:25.633	0:04.170
B9_EJ_250509	0:30.442	0:32.009	0:01.566
A10_EJ_250509	2:11.564	2:14.834	0:03.269
A9_EJ_250509	2:34.615	2:36.886	0:02.271
A7_EJ_250509	2:59.074	3:00.611	0:01.536
A4_EJ_250509	3:07.104	3:08.929	0:01.824
A2_EJ_250509	3:10.898	3:13.222	0:02.324

Note: File start with B= Before; Filestart with A= After

T R A C K L I S T I N G Total: 20

MIP Session: Louise_Cliona

TRACK NAME:	Louise		
FILE NAME	START TIME	END TIME	DURATION
B2_LC_260309	0:28.940	0:36.844	0:07.903
B4_LC_260309	0:40.005	0:44.505	0:04.499
B5_LC_260309	0:44.930	0:48.821	0:03.891
B6_LC_260309	0:49.612	0:51.679	0:02.067
B8_LC_260309	0:54.962	0:58.002	0:03.039
B10_LC_260309	1:04.325	1:07.365	0:03.039
A9_LC_260309	10:25.990	10:28.239	0:02.249
A7_LC_260309	10:34.502	10:37.177	0:02.675
A6_LC_260309	10:38.271	10:40.947	0:02.675
A4_LC_260309	10:54.687	10:58.457	0:03.769
A3_LC_260309	11:01.740	11:03.382	0:01.641
A1_LC_260309	11:03.625	11:08.063	0:04.438

TRACK NAME:	Cliona		
FILE NAME	START TIME	END TIME	DURATION
B1_LC_260309	0:25.961	0:29.366	0:03.404
B3_LC_260309	0:37.087	0:41.404	0:04.316
B7_LC_260309	0:52.773	0:56.178	0:03.404
B9_LC_260309	0:56.239	0:58.975	0:02.735
A10_LC_260309	10:10.243	10:11.763	0:01.519
A8_LC_260309	10:30.854	10:35.475	0:04.620
A5_LC_260309	10:46.479	10:48.790	0:02.310
A2_LC_260309	11:02.652	11:06.725	0:04.073

T R A C K L I S T I N G Total: 20

MIP Session: Jack_Keith

TRACK NAME:	Jack		
FILE NAME	START TIME	END TIME	DURATION
B1_JK_210809	0:08.889	0:10.565	0:01.675
B2_JK_210809	0:10.871	0:12.281	0:01.410
B3_JK_210809	0:13.037	0:14.345	0:01.307
B4_JK_210809	0:21.913	0:26.111	0:04.198
B5_JK_210809	0:26.116	0:28.446	0:02.329
B7_JK_210809	1:07.152	1:08.822	0:01.669
B8_JK_210809	1:14.734	1:17.404	0:02.670
A10_JK_210809	8:07.926	8:11.305	0:03.379
A9_JK_210809	8:13.456	8:16.630	0:03.174
A7_JK_210809	8:39.848	8:43.470	0:03.622
A5_JK_210809	9:41.149	9:43.471	0:02.322
A3_JK_210809	10:12.819	10:15.100	0:02.280
A2_JK_210809	10:24.222	10:26.566	0:02.343
A1_JK_210809	11:01.954	11:05.112	0:03.157

TRACK NAME:	Keith		
FILE NAME	START TIME	END TIME	DURATION
B6_JK_210809	0:32.819	0:35.006	0:02.186
B9_JK_210809	1:38.569	1:44.059	0:05.489
B10_JK_210809	1:51.204	1:55.197	0:03.993
A8_JK_210809	8:16.814	8:20.436	0:03.622
A6_JK_210809	9:10.870	9:13.656	0:02.786
A4_JK_210809	10:03.634	10:06.358	0:02.724

Note: File start with B= Before; Filestart with A= After

T R A C K L I S T I N G Total: 20

MIP Session:Jenny_Ruth

TRACK NAME:	Jenny		
FILE NAME	START TIME	END TIME	DURATION
B2_JR_240309	0:11.098	0:13.390	0:02.291
B3_JR_240309	0:14.402	0:17.880	0:03.477
B5_JR_240309	0:21.256	0:24.678	0:03.421
B6_JR_240309	0:25.362	0:27.053	0:01.690
B7_JR_240309	0:30.756	0:32.267	0:01.510
B8_JR_240309	0:33.559	0:36.144	0:02.584
B10_JR_240309	0:37.561	0:39.979	0:02.417
A9_JR_240309	8:36.294	8:39.006	0:02.711
A7_JR_240309	9:46.934	9:48.643	0:01.709
A6_JR_240309	9:54.750	9:57.293	0:02.543
A4_JR_240309	10:05.131	10:07.153	0:02.021
A1_JR_240309	10:08.779	10:12.322	0:03.543

TRACK NAME:	Ruth		
FILE NAME	START TIME	END TIME	DURATION
B1_JR_240309	0:08.615	0:11.272	0:02.657
B4_JR_240309	0:18.599	0:21.538	0:02.938
B9_JR_240309	0:36.394	0:38.561	0:02.167
A10_JR_240309	8:25.267	8:26.970	0:01.702
A8_JR_240309	9:05.689	9:09.409	0:03.719
A5_JR_240309	10:00.983	10:03.484	0:02.501
A3_JR_240309	10:05.047	10:08.028	0:02.980
A2_JR_240309	10:09.675	10:10.842	0:01.167

T R A C K L I S T I N G Total: 20

MIP Session: Paula_Anna

TRACK NAME: Paul

FILE NAME	START TIME	END TIME	DURATION
B2_PA_090609	0:09.102	0:12.074	0:02.972
B5_PA_090609	0:40.867	0:46.439	0:05.572
B6_PA_090609	1:00.743	1:07.245	0:06.501
B7_PA_090609	1:13.560	1:18.019	0:04.458
B10_PA_090609	1:22.291	1:27.678	0:05.387
A9_PA_090609	7:21.237	7:24.032	0:02.795
A8_PA_090609	7:35.482	7:39.012	0:03.529
A6_PA_090609	8:17.122	8:20.993	0:03.870

TRACK NAME: Anna

FILE NAME	START TIME	END TIME	DURATION
B1_PA_090609	0:04.086	0:08.916	0:04.829
B3_PA_090609	0:09.845	0:13.003	0:03.157
B4_PA_090609	0:39.381	0:44.768	0:05.387
B8_PA_090609	1:15.418	1:18.762	0:03.343
B9_PA_090609	1:19.133	1:22.477	0:03.343
A10_PA_090609	7:19.321	7:23.965	0:04.644
A7_PA_090609	7:47.371	7:51.829	0:04.458
A5_PA_090609	8:39.569	8:43.656	0:04.086
A4_PA_090609	9:20.808	9:27.681	0:06.873
A3_PA_090609	10:14.678	10:17.836	0:03.157
A2_PA_090609	10:23.223	10:29.725	0:06.501
A1_PA_090609	10:48.858	10:54.988	0:06.130

Note: File start with B= Before; Filestart with A= After

T R A C K L I S T I N G Total: 20

MIP Session: Pauly_John

TRACK NAME: John

FILE NAME	START TIME	END TIME	DURATION
B3_PJ_090609	0:11.600	0:12.843	0:01.242
B4_PJ_090609	0:22.467	0:24.729	0:02.262
B6_PJ_090609	0:30.026	0:31.295	0:01.269
B9_PJ_090609	0:35.798	0:39.042	0:03.244
B10_PJ_090609	0:39.042	0:42.485	0:03.443
A10_PJ_090609	9:00.770	9:04.566	0:03.796
A9_PJ_090609	9:06.817	9:10.481	0:03.663
A7_PJ_090609	9:19.750	9:21.428	0:01.677
A5_PJ_090609	9:26.813	9:29.726	0:02.913
A3_PJ_090609	9:32.198	9:34.891	0:02.692
A2_PJ_090609	9:42.836	9:45.529	0:02.692

TRACK NAME: Pauly

FILE NAME	START TIME	END TIME	DURATION
B1_PJ_090609	0:01.037	0:03.200	0:02.162
B2_PJ_090609	0:11.167	0:13.816	0:02.648
B5_PJ_090609	0:29.066	0:30.214	0:01.147
B7_PJ_090609	0:31.251	0:32.642	0:01.390
B8_PJ_090609	0:32.995	0:35.974	0:02.979
A8_PJ_090609	9:11.937	9:15.425	0:03.487
A6_PJ_090609	9:24.032	9:27.210	0:03.178
A4_PJ_090609	9:28.711	9:31.359	0:02.648
A1_PJ_090609	9:44.293	9:45.396	0:01.103

D

Ratings Summary

Summary of obtained values for all clips

	Evaluation												Activation											
Clip	M	Mdn	Mode	GT	SD	VAR	Range	Min	Max	25th	50th	IQR	M	Mdn	Mode	GT	SD	VAR	Range	Min	Max	25th	75th	IQR
A1_CC_200309	3.00	3	3	2.932	0.53	0.29	2	2	4	3	3	0	3.38	3.5	4	3.369	0.74	0.55	2	2	4	3	4	1
A1_DL_230309	2.23	2	4	1.969	1.79	3.19	4	0	4	0.5	4	3.5	4.00	4	4	4	0.00	0.00	0	4	4	4	4	0
A1_EJ_250509	1.80	1.5	1	2	1.23	1.51	4	0	4	1	3	2	3.40	4	4	3.235	1.07	1.16	3	1	4	2.75	4	1.25
A1_JK_210809	1.50	1	1	1	1.02	1.04	4	0	4	1	2	1	2.71	3	2a	2.449	0.73	0.53	2	2	4	2	3	1
A1_JR_240309	0.92	1	1	1	0.76	0.58	2	0	2	0	1.5	1.5	3.31	3	3a	3.249	0.85	0.73	3	1	4	3	4	1
A1_LC_260309	0.86	1	0	0.683	0.90	0.81	2	0	2	0	2	2	2.57	3	4	3.017	1.81	3.29	4	0	4	0	4	4
A1_PA_090609	2.15	2	3	3	1.34	1.81	4	0	4	1	3	2	1.92	2	2	2.3	1.32	1.74	4	0	4	1	3	2
A1_PJ_090609	1.21	1	1	1.195	0.80	0.64	3	0	3	1	2	1	2.50	3	3	2.2	0.85	0.73	3	1	4	2	3	1
A10_CC_200309	2.22	2	2	2	0.67	0.44	2	2	4	2	2	0	2.22	2	2	2.126	0.44	0.19	1	2	3	2	2.5	0.5
A10_DL_230309	3.77	4	4	4	0.44	0.19	1	3	4	3.5	4	0.5	3.92	4	4	3.922	0.28	0.08	1	3	4	4	4	0
A10_EJ_250509	2.10	2	2	1.946	0.57	0.32	2	1	3	2	2.25	0.25	1.00	1	1	1.08	0.67	0.44	2	0	2	0.75	1.25	0.5
A10_JK_210809	2.00	2	2	2	0.67	0.44	2	1	3	1.75	2.25	0.5	2.60	2.5	2	3	0.70	0.49	2	2	4	2	3	1
A10_JR_240309	0.90	0.5	0	0.762	1.29	1.66	4	0	4	0	1.25	1.25	3.70	4	4	3.627	0.48	0.23	1	3	4	3	4	1
A10_LC_260309	1.33	1	1	1.517	1.03	1.07	3	0	3	0.75	2.25	1.5	2.83	3	3	2.601	0.75	0.57	2	2	4	2	3.25	1.25
A10_PA_090609	2.40	3	3	2.493	0.97	0.93	2	1	3	1	3	2	2.30	3	3	2.886	1.42	2.01	4	0	4	0.75	3	2.25
A10_PJ_090609	2.18	2	2	2.098	0.40	0.16	1	2	3	2	2	0	2.09	2	2a	1.802	0.83	0.69	2	1	3	1	3	2
A2_CC_200309	1.00	1	0a	0.574	1.22	1.50	3	0	3	0	2	2	1.80	2	3	1.099	1.30	1.70	3	0	3	0.5	3	2.5
A2_DL_230309	2.77	3	3	2.763	1.30	1.69	4	0	4	2	4	2	3.62	4	4	3.807	0.65	0.42	2	2	4	3	4	1
A2_EJ_250509	1.63	2	2	1.819	0.52	0.27	1	1	2	1	2	1	1.88	2	2	1.87	0.64	0.41	2	1	3	1.25	2	0.75
A2_JK_210809	1.36	1	1a	1.457	0.67	0.45	2	0	2	1	2	1	2.36	2	2	2.479	1.21	1.45	4	0	4	2	3	1
A2_JR_240309	1.75	2	2	2	1.48	2.20	4	0	4	0	2.75	2.75	3.58	4	4	3.58	0.51	0.27	1	3	4	3	4	1
A2_LC_260309	2.00	2	1a	1.975	1.55	2.40	4	0	4	0.75	3.25	2.5	1.83	2	2	1.497	0.75	0.57	2	1	3	1	2.25	1.25
A2_PA_090609	1.60	2	2	2	0.70	0.49	2	0	2	1	2	1	1.30	1	1	1.339	0.95	0.90	3	0	3	0.75	2	1.25
A2_PJ_090609	2.55	3	1a	2	1.37	1.87	3	1	4	1	4	3	3.00	3	3	3.013	0.77	0.60	3	1	4	3	3	0
A3_CC_200309	1.18	1	0	0	1.47	2.16	4	0	4	0	3	3	2.09	2	3	1.834	0.94	0.89	2	1	3	1	3	2
A3_DL_230309	2.00	2	0a	1.697	1.76	3.09	4	0	4	0	4	4	3.67	4	4	3.763	0.49	0.24	1	3	4	3	4	1
A3_EJ_250509	2.00	2	1a	2.105	1.00	1.00	2	1	3	1	3	2	2.00	2	1a	1.679	1.00	1.00	2	1	3	1	3	2
A3_JK_210809	1.38	1	1	1	0.96	0.92	4	0	4	1	2	1	2.38	2	2	2.111	1.12	1.26	3	1	4	1.5	3.5	2
A3_JR_240309	2.71	4	4	2.945	1.89	3.57	4	0	4	0	4	4	4.00	4	4	4	0.00	0.00	0	4	4	4	4	0
A3_LC_260309	0.92	1	1	2	0.79	0.63	2	0	2	0	1.75	1.75	0.83	0	0	0.639	1.27	1.61	3	0	3	0	2	2
A3_PA_090609	2.67	3	4	2	1.22	1.50	3	1	4	1.5	4	2.5	3.11	3	3	3.202	0.93	0.86	3	1	4	3	4	1
A3_PJ_090609	1.67	2	2	2	0.50	0.25	1	1	2	1	2	1	0.78	1	1	0.918	0.67	0.44	2	0	2	0	1	1
A4_CC_200309	1.75	2	2	1.775	1.06	1.11	3	0	3	1	2.75	1.75	1.25	1.5	2	1.45	1.06	1.11	3	0	3	0	2	2
A4_DL_230309	2.09	2	1	2	1.04	1.09	3	1	4	1	3	2	3.36	3	3a	3.428	0.67	0.45	2	2	4	3	4	1
A4_EJ_250509	1.36	1	1a	2	0.67	0.45	2	0	2	1	2	1	1.55	1	1	1.527	1.13	1.27	3	0	3	1	3	2
A4_JK_210809	2.33	2	2a	2.102	1.00	1.00	3	1	4	1.5	3	1.5	2.11	2	2	2.016	0.78	0.61	2	1	3	1.5	3	1.5
A4_JR_240309	2.78	4	4	2.406	1.86	3.44	4	0	4	0.5	4	3.5	3.89	4	4	3.888	0.33	0.11	1	3	4	4	4	0
A4_LC_260309	3.18	4	4	4	1.25	1.56	4	0	4	3	4	1	3.82	4	4	3.734	0.40	0.16	1	3	4	4	4	0
A4_PA_090609	3.10	3	3	3	0.74	0.54	2	2	4	2.75	4	1.25	2.30	3	3	2.641	1.16	1.34	3	1	4	1	3	2
A4_PJ_090609	2.13	2	1a	2	1.25	1.55	3	1	4	1	3.5	2.5	3.25	3	3	3.092	0.71	0.50	2	2	4	3	4	1
A5_CC_200309	1.00	1	0	1.196	1.15	1.33	3	0	3	0	2	2	0.86	1	0	0.73	0.90	0.81	2	0	2	0	2	2

M = Mean; Mdn = Median; Mo = Mode; SD = Standard Deviation; R = Range; 25th = First quartile; 75th = Third quartile; IQR = Interquartile Range.

a. Multiple modes exist. The smallest value is shown

Summary of obtained values for all clips

	Evaluation												Activation											
Asset	M	Mdn	Mode	GT	SD	VAR	Range	Min	Max	25th	50th	IQR	M	Mdn	Mode	GT	SD	VAR	Range	Min	Max	25th	75th	IQR
A5_DL_230309	2.50	3	3	2.707	1.18	1.39	3	1	4	1	3.25	2.25	3.20	3	3	3.209	0.42	0.18	1	3	4	3	3.25	0.25
A5_EJ_250509	2.67	3	3	3	1.12	1.25	3	1	4	1.5	3.5	2	3.44	3	3	3.375	0.53	0.28	1	3	4	3	4	1
A5_JK_210809	1.56	1	1	1.546	1.01	1.03	3	0	3	1	2.5	1.5	3.00	3	4	3.099	1.32	1.75	4	0	4	2.5	4	1.5
A5_JR_240309	2.79	3	3	3	1.05	1.10	3	1	4	2	4	2	3.36	4	4	3.538	1.15	1.32	4	0	4	3	4	1
A5_LC_260309	1.40	1.5	2	2	0.70	0.49	2	0	2	1	2	1	2.10	2	2	2.115	0.88	0.77	3	0	3	2	3	1
A5_PA_090609	2.70	3	4	3.106	1.42	2.01	4	0	4	1.75	4	2.25	2.90	3	3	3.307	1.10	1.21	3	1	4	2.5	4	1.5
A5_PJ_090609	1.71	2	2	1.867	0.61	0.37	2	0	2	1.75	2	0.25	1.79	2	1	2.08	0.97	0.95	3	0	3	1	3	2
A6_CC_200309	2.08	2	2	1.812	1.04	1.08	3	1	4	1	2.5	1.5	2.85	3	3	2.803	0.99	0.97	3	1	4	2.5	3.5	1
A6_DL_230309	3.00	3.5	4	3	1.36	1.85	4	0	4	2.5	4	1.5	3.86	4	4	3.807	0.36	0.13	1	3	4	4	4	0
A6_EJ_250509	1.50	1	1	1.257	1.45	2.12	4	0	4	0.75	2.5	1.75	3.64	4	4	3.788	0.84	0.71	3	1	4	3.75	4	0.25
A6_JK_210809	1.73	2	2	1.978	1.19	1.42	4	0	4	1	2	1	3.09	3	3	2.805	0.70	0.49	2	2	4	3	4	1
A6_JR_240309	1.82	1	1	2	1.47	2.16	4	0	4	1	3	2	3.45	4	4	3.396	0.69	0.47	2	2	4	3	4	1
A6_LC_260309	1.67	2	2	2	0.65	0.42	2	0	2	1.25	2	0.75	1.92	2	2a	1.855	1.00	0.99	3	0	3	1	3	2
A6_PA_090609	2.50	2	2a	2.094	1.31	1.71	3	1	4	1.25	4	2.75	2.88	3	3	2.47	0.83	0.70	3	1	4	3	3	0
A6_PJ_090609	2.33	2	2	2.12	0.71	0.50	2	2	4	2	2.5	0.5	2.67	3	2a	2.626	0.71	0.50	2	2	4	2	3	1
A7_CC_200309	1.50	2	2	1.626	0.65	0.42	2	0	2	1	2	1	2.64	3	3	2.542	0.84	0.71	3	1	4	2	3	1
A7_DL_230309	2.47	3	3	2	0.83	0.70	3	1	4	2	3	1	2.60	3	3	2.481	0.74	0.54	3	1	4	2	3	1
A7_EJ_250509	2.00	2	2	2.125	0.95	0.91	3	0	3	1.25	3	1.75	1.92	2	2	2.014	0.79	0.63	3	0	3	2	2	0
A7_JK_210809	2.10	2	2	1.836	1.10	1.21	3	1	4	1	2.5	1.5	3.10	3	3	2.842	0.74	0.54	2	2	4	2.75	4	1.25
A7_JR_240309	1.00	1	0	0.736	1.22	1.50	3	0	3	0	2	2	3.89	4	4	3.903	0.33	0.11	1	3	4	4	4	0
A7_LC_260309	1.27	1	1	1	0.79	0.62	3	0	3	1	2	1	1.73	2	2	1.99	0.90	0.82	3	0	3	1	2	1
A7_PA_090609	1.18	1	1	1.452	0.87	0.76	3	0	3	1	2	1	1.64	2	1a	1.722	1.12	1.25	3	0	3	1	3	2
A7_PJ_090609	1.83	2	2	2	0.83	0.70	3	0	3	1.25	2	0.75	2.33	2.5	3	2.216	1.07	1.15	4	0	4	2	3	1
A8_CC_200309	2.38	2	2	2.3	0.74	0.55	2	2	4	2	2.75	0.75	2.63	2.5	2	2.51	0.74	0.55	2	2	4	2	3	1
A8_DL_230309	2.50	2.5	2a	2.204	0.93	0.86	3	1	4	2	3	1	3.13	3.5	4	3.327	1.13	1.27	3	1	4	2.25	4	1.75
A8_EJ_250509	1.67	2	2	2	0.65	0.42	2	1	3	1	2	1	2.92	3	3	2.885	0.79	0.63	2	2	4	2	3.75	1.75
A8_JK_210809	0.71	1	0a	0.724	0.76	0.57	2	0	2	0	1	1	1.29	1	1	1.143	0.95	0.90	3	0	3	1	2	1
A8_JR_240309	1.22	0	0	1.156	1.86	3.44	4	0	4	0	3.5	3.5	4.00	4	4	4	0.00	0.00	0	4	4	4	4	0
A8_LC_260309	1.82	2	2	1.782	0.60	0.36	2	1	3	1	2	1	1.82	2	3	1.992	1.25	1.56	3	0	3	1	3	2
A8_PA_090609	2.07	2	3	2.06	1.16	1.35	4	0	4	1	3	2	2.00	2	3	2.252	1.07	1.14	3	0	3	1	3	2
A8_PJ_090609	2.55	2	2	2	0.82	0.67	2	2	4	2	3	1	2.64	2	2	2	0.81	0.65	2	2	4	2	3	1
A9_CC_200309	1.55	2	2	2	0.69	0.47	2	0	2	1	2	1	1.09	1	1	1.091	0.83	0.69	3	0	3	1	1	0
A9_DL_230309	2.67	3	3a	3	1.40	1.95	4	0	4	2	4	2	3.73	4	4	3.608	0.46	0.21	1	3	4	3	4	1
A9_EJ_250509	2.00	2	2	1.985	0.53	0.29	2	1	3	2	2	0	2.75	2.5	2	2.608	0.89	0.79	2	2	4	2	3.75	1.75
A9_JK_210809	2.14	2	2	2	0.95	0.90	4	0	4	2	3	1	2.43	3	3	2.283	1.16	1.34	4	0	4	1.75	3	1.25
A9_JR_240309	1.70	1.5	0	1.315	1.70	2.90	4	0	4	0	3.25	3.25	3.90	4	4	3.895	0.32	0.10	1	3	4	4	4	0
A9_LC_260309	2.44	2	2	2	1.01	1.03	3	1	4	2	3.5	1.5	2.44	3	3	2.398	1.24	1.53	4	0	4	1.5	3	1.5
A9_PA_090609	1.91	2	2	2	0.83	0.69	3	1	4	1	2	1	1.55	2	2	1.487	1.13	1.27	4	0	4	1	2	1
A9_PJ_090609	1.45	1	1	1	0.93	0.87	3	0	3	1	2	1	0.18	0	0	0.091	0.40	0.16	1	0	1	0	0	0
B1_CC_200309	2.58	3	3	2.371	1.08	1.17	3	1	4	1.25	3	1.75	2.25	2	2a	2.108	0.97	0.93	3	1	4	1.25	3	1.75
B1_DL_230309	2.60	2.5	2	2.576	0.70	0.49	2	2	4	2	3	1	1.40	1	0a	1.573	1.35	1.82	4	0	4	0	2.25	2.25

M = Mean; Mdn = Median; Mo = Mode; SD = Standard Deviation; R = Range; 25th = First quartile; 75th = Third quartile; IQR = Interquartile Range.
a. Multiple modes exist. The smallest value is shown

Summary of obtained values for all clips

	Evaluation												Activation											
Asset	M	Mdn	Mode	GT	SD	VAR	Range	Min	Max	25th	50th	IQR	M	Mdn	Mode	GT	SD	VAR	Range	Min	Max	25th	75th	IQR
B1_EJ_250509	1.62	2	2	2	0.96	0.92	3	0	3	1	2	1	2.08	2	2	2.186	0.86	0.74	3	0	3	2	3	1
B1_JK_210809	2.92	3	3a	2.719	1.00	0.99	3	1	4	2	4	2	3.67	4	4	3.624	0.49	0.24	1	3	4	3	4	1
B1_JR_240309	1.91	2	2	1.932	0.70	0.49	2	1	3	1	2	1	2.09	2	3	2.148	1.22	1.49	4	0	4	1	3	2
B1_LC_260309	2.29	2	2	2	0.61	0.37	2	2	4	2	2.25	0.25	2.07	2	2	2.077	1.07	1.15	4	0	4	1	3	2
B1_PA_090609	2.67	3	3	2.734	0.87	0.75	3	1	4	2	3	1	2.67	3	3	2.801	0.87	0.75	3	1	4	2	3	1
B1_PJ_090609	1.80	2	2	2	0.92	0.84	3	0	3	1	2.25	1.25	2.20	3	3	2.659	1.62	2.62	4	0	4	0	3.25	3.25
B10_CC_200309	1.43	2	2	1	0.94	0.88	3	0	3	0.75	2	1.25	0.57	0	0	0.404	0.85	0.73	3	0	3	0	1	1
B10_DL_230309	2.08	2	2	2	0.79	0.63	3	1	4	2	2	0	2.08	2	2	1.858	1.24	1.54	4	0	4	1	3	2
B10_EJ_250509	1.40	1	1	2	0.52	0.27	1	1	2	1	2	1	2.00	2	3	2.198	1.05	1.11	3	0	3	1	3	2
B10_JK_210809	1.50	1	1	1.5	1.27	1.61	4	0	4	0.75	2.25	1.5	2.60	3	3	2.397	1.07	1.16	4	0	4	2	3	1
B10_JR_240309	1.38	1.5	2	1.55	1.06	1.13	3	0	3	0.25	2	1.75	2.63	3	3	2.638	1.30	1.70	4	0	4	2	3.75	1.75
B10_LC_260309	2.25	2	2	2.213	0.46	0.21	1	2	3	2	2.75	0.75	2.50	2	2	2.298	0.76	0.57	2	2	4	2	3	1
B10_PA_090609	1.63	1	1	1.553	1.30	1.70	4	0	4	1	2.75	1.75	2.38	2.5	1	1.95	1.30	1.70	3	1	4	1	3.75	2.75
B10_PJ_090609	2.00	2	2	1.931	0.58	0.33	2	1	3	2	2	0	1.57	2	2	1.625	0.53	0.29	1	1	2	1	2	1
B2_CC_200309	1.07	1	1	1	0.83	0.69	3	0	3	0.75	1.25	0.5	1.29	1	0	1.329	1.27	1.60	3	0	3	0	3	3
B2_DL_230309	2.15	2	3	2.179	1.34	1.81	4	0	4	1	3	2	2.23	2	3	2.131	1.01	1.03	3	1	4	1	3	2
B2_EJ_250509	1.82	2	2	1.624	0.60	0.36	2	1	3	1	2	1	2.36	2	2	2.175	1.03	1.05	3	1	4	2	3	1
B2_JK_210809	2.13	2	2	2	0.83	0.70	3	1	4	2	2	0	3.13	3	3	2.888	0.64	0.41	2	2	4	3	3.75	0.75
B2_JR_240309	1.71	2	1a	1.728	0.76	0.57	2	1	3	1	2	1	2.00	2	1a	1.837	1.53	2.33	4	0	4	1	4	3
B2_LC_260309	2.20	2.5	3	1	0.92	0.84	2	1	3	1	3	2	2.90	3	3	2.644	0.88	0.77	3	1	4	2.75	3.25	0.5
B2_PA_090609	1.80	2	2	1.847	1.14	1.29	4	0	4	1	2.25	1.25	2.00	2	2	1.932	1.05	1.11	4	0	4	1.75	2.25	0.5
B2_PJ_090609	2.36	2	2	2.256	1.01	1.02	4	0	4	2	3	1	2.64	3	2a	2.588	1.08	1.17	4	0	4	2	3.25	1.25
B3_CC_200309	2.08	2	2	1.935	0.76	0.58	3	1	4	2	2	0	1.54	2	2	1.497	0.97	0.94	3	0	3	1	2	1
B3_DL_230309	1.91	2	2	2	0.54	0.29	2	1	3	2	2	0	1.64	2	3	1.863	1.29	1.65	3	0	3	0	3	3
B3_EJ_250509	1.85	2	2	1.994	0.99	0.97	3	0	3	1.5	2.5	1	2.69	3	3	2.613	0.95	0.90	3	1	4	2	3	1
B3_JK_210809	2.57	2.5	2	2	0.85	0.73	3	1	4	2	3	1	2.93	3	2a	2.695	0.83	0.69	2	2	4	2	4	2
B3_JR_240309	1.10	1	1	1	0.74	0.54	2	0	2	0.75	2	1.25	2.90	3	3	2.662	0.74	0.54	2	2	4	2	3.25	1.25
B3_LC_260309	1.36	1	1	1.431	0.74	0.55	3	0	3	1	2	1	3.00	3	3	3.1	1.04	1.08	4	0	4	3	4	1
B3_PA_090609	1.73	2	2	2	0.47	0.22	1	1	2	1	2	1	1.36	2	2	1.574	1.03	1.05	3	0	3	0	2	2
B3_PJ_090609	1.65	2	1	2	0.70	0.49	2	1	3	1	2	1	1.29	1	1	1.227	0.92	0.85	3	0	3	1	2	1
B4_CC_200309	3.56	4	4	3.534	0.53	0.28	1	3	4	3	4	1	3.44	4	4	3.353	0.73	0.53	2	2	4	3	4	1
B4_DL_230309	1.57	2	2	2	0.53	0.29	1	1	2	1	2	1	2.14	2	3	2.103	0.90	0.81	2	1	3	1	3	2
B4_EJ_250509	2.09	2	2	2.031	0.70	0.49	2	1	3	2	3	1	1.82	2	3	1.805	1.33	1.76	3	0	3	0	3	3
B4_JK_210809	1.67	2	2	1	0.62	0.38	2	1	3	1	2	1	1.60	2	1a	1.507	1.12	1.26	3	0	3	1	3	2
B4_JR_240309	1.71	2	2	2	1.07	1.14	4	0	4	1	2	1	2.57	3	3	2.33	0.94	0.88	3	1	4	2	3	1
B4_LC_260309	2.20	2	2	2	0.92	0.84	3	0	3	2	3	1	2.50	2.5	2a	2.356	0.85	0.72	3	1	4	2	3	1
B4_PA_090609	3.17	3	3a	3	0.79	0.62	2	2	4	2.75	4	1.25	3.11	3	3a	3	0.90	0.81	3	1	4	2.75	4	1.25
B4_PJ_090609	2.00	2	2	1.944	1.00	1.00	3	0	3	2	3	1	1.29	1	1a	1.461	0.76	0.57	2	0	2	1	2	1
B5_CC_200309	2.70	2.5	2	2.507	0.82	0.68	2	2	4	2	3.25	1.25	2.80	3	2a	2.661	0.79	0.62	2	2	4	2	3.25	1.25
B5_DL_230309	1.56	1	1	1.497	0.73	0.53	2	1	3	1	2	1	1.89	2	2a	1.739	1.05	1.11	3	0	3	1	3	2
B5_EJ_250509	1.50	2	2	1	0.71	0.50	2	0	2	1	2	1	1.60	1.5	1	1.734	0.70	0.49	2	1	3	1	2	1

M = Mean; Mdn = Median; Mo = Mode; SD = Standard Deviation; R = Range; 25th = First quartile; 75th = Third quartile; IQR = Interquartile Range.
a. Multiple modes exist. The smallest value is shown

Summary of obtained values for all clips

	Evaluation												Activation											
Asset	M	Mdn	Mode	GT	SD	VAR	Range	Min	Max	25th	50th	IQR	M	Mdn	Mode	GT	SD	VAR	Range	Min	Max	25th	75th	IQR
B5_JK_210809	2.08	2	2	2.069	0.86	0.74	3	1	4	1.5	2.5	1	2.46	3	3	2.397	1.05	1.10	4	0	4	2	3	1
B5_JR_240309	2.31	2	1a	2.248	1.11	1.23	3	1	4	1	3	2	2.85	3	3	2.853	0.99	0.97	3	1	4	2.5	3.5	1
B5_LC_260309	2.40	3	3	3	0.97	0.93	3	0	3	2	3	1	2.50	3	3	2.382	0.71	0.50	2	1	3	2	3	1
B5_PA_090609	3.50	4	4	3.548	0.93	0.86	2	2	4	2.5	4	1.5	3.75	4	4	3.799	0.46	0.21	1	3	4	3.25	4	0.75
B5_PJ_090609	1.70	2	2	1.63	1.25	1.57	4	0	4	0.75	2.25	1.5	3.30	3	3	3.153	0.67	0.46	2	2	4	3	4	1
B6_CC_200309	2.64	2	2	4	1.03	1.05	3	1	4	2	4	2	1.82	2	1a	1.82	1.40	1.96	4	0	4	1	3	2
B6_DL_230309	2.00	2	2	2	1.04	1.09	4	0	4	1.25	2.75	1.5	2.17	2	2a	2.076	1.11	1.24	4	0	4	1.25	3	1.75
B6_EJ_250509	1.86	2	2	2	0.53	0.29	2	1	3	1.75	2	0.25	1.00	1	1	1.032	0.88	0.77	3	0	3	0	1.25	1.25
B6_JK_210809	2.50	2	2	2.353	0.84	0.70	2	2	4	2	3.25	1.25	2.50	3	3	2.546	1.38	1.90	4	0	4	1.5	3.25	1.75
B6_JR_240309	1.58	1.5	1a	1	1.16	1.36	4	0	4	1	2	1	1.83	2	2	1.642	0.94	0.88	3	0	3	1	2.75	1.75
B6_LC_260309	1.38	1.5	2	1.485	0.74	0.55	2	0	2	1	2	1	1.13	1	0	1.361	1.13	1.27	3	0	3	0	2	2
B6_PA_090609	2.40	2	2	2	0.84	0.71	3	1	4	2	3	1	1.30	1	1	1.438	0.82	0.68	3	0	3	1	2	1
B6_PJ_090609	1.92	2	2	2	0.28	0.08	1	1	2	2	2	0	1.62	2	1a	1.802	0.87	0.76	3	0	3	1	2	1
B7_CC_200309	1.78	2	2	2	0.67	0.44	2	1	3	1	2	1	1.89	1	1	1.521	1.17	1.36	3	1	4	1	3	2
B7_DL_230309	2.25	3	3	3	1.06	1.11	3	0	3	1.25	3	1.75	1.42	1	1	1	0.90	0.81	3	0	3	1	2	1
B7_EJ_250509	1.38	1.5	2	1.493	1.06	1.13	3	0	3	0.25	2	1.75	0.50	0	0	0.142	1.07	1.14	3	0	3	0	0.75	0.75
B7_JK_210809	2.50	2.5	2a	2.255	1.08	1.17	3	1	4	1.75	3.25	1.5	3.20	3	3	3.087	0.63	0.40	2	2	4	3	4	1
B7_JR_240309	1.50	1.5	1a	1.479	0.52	0.27	1	1	2	1	2	1	1.83	2	2	1.758	0.94	0.88	3	0	3	1	2.75	1.75
B7_LC_260309	1.69	1	1	1.591	1.18	1.40	4	0	4	1	3	2	1.69	1	1a	1.3	1.38	1.90	4	0	4	0.5	3	2.5
B7_PA_090609	1.33	2	2	1.641	1.00	1.00	2	0	2	0	2	2	2.22	2	2	2.156	1.09	1.19	4	0	4	2	3	1
B7_PJ_090609	2.10	2	2	2.026	0.57	0.32	2	1	3	2	2.25	0.25	2.40	3	3	2.477	0.97	0.93	3	0	3	2	3	1
B8_CC_200309	1.64	2	2	1.51	0.50	0.25	1	1	2	1	2	1	1.64	2	2	1.722	0.93	0.86	3	0	3	1	2	1
B8_DL_230309	2.50	2.5	2a	2.357	0.93	0.86	3	1	4	2	3	1	3.25	3	3	3.023	0.71	0.50	2	2	4	3	4	1
B8_EJ_250509	2.38	2	2	2	0.87	0.76	3	1	4	2	3	1	2.69	3	3	2.596	1.11	1.23	4	0	4	2	3.5	1.5
B8_JK_210809	1.57	2	2	2	0.53	0.29	1	1	2	1	2	1	1.86	2	2	1.947	0.69	0.48	2	1	3	1	2	1
B8_JR_240309	2.00	2	2	2	0.50	0.25	2	1	3	2	2	0	2.11	2	2	2.066	0.78	0.61	2	1	3	1.5	3	1.5
B8_LC_260309	3.00	3	4	3	1.18	1.40	3	1	4	2	4	2	2.91	3	3	3.19	1.30	1.69	4	0	4	3	4	1
B8_PA_090609	2.27	2	2	2.247	0.47	0.22	1	2	3	2	3	1	2.64	3	3	2.577	0.50	0.25	1	2	3	2	3	1
B8_PJ_090609	2.33	2	2	2	0.71	0.50	2	2	4	2	2.5	0.5	2.78	3	3	2.602	0.97	0.94	3	1	4	2	3.5	1.5
B9_CC_200309	1.63	2	2	1.624	0.52	0.27	1	1	2	1	2	1	2.50	2.5	2a	2.442	0.53	0.29	1	2	3	2	3	1
B9_DL_230309	2.36	2	2	2	0.81	0.65	3	1	4	2	3	1	3.18	3	3	3.033	0.60	0.36	2	2	4	3	4	1
B9_EJ_250509	1.86	2	2	2	0.38	0.14	1	1	2	2	2	0	0.71	0	0	0.664	0.95	0.90	2	0	2	0	2	2
B9_JK_210809	2.88	3	3	2.515	0.99	0.98	3	1	4	2.25	3.75	1.5	3.00	3	3	2.765	0.76	0.57	2	2	4	2.25	3.75	1.5
B9_JR_240309	1.13	1	1	0.836	0.99	0.98	3	0	3	0.25	1.75	1.5	1.63	1.5	0a	1.572	1.41	1.98	4	0	4	0.25	2.75	2.5
B9_LC_260309	3.00	4	4	3.149	1.32	1.75	3	1	4	1.5	4	2.5	3.00	3	4	3.096	1.22	1.50	3	1	4	2	4	2
B9_PA_090609	2.19	2	2	2	0.54	0.30	2	1	3	2	2.75	0.75	1.50	1	1	1.423	0.82	0.67	3	0	3	1	2	1
B9_PJ_090609	1.58	2	2	1.612	0.79	0.63	3	0	3	1	2	1	1.08	1	1	1.169	0.67	0.45	2	0	2	1	1.75	0.75

M = Mean; Mdn = Median; Mo = Mode; SD = Standard Deviation; R = Range; 25th = First quartile; 75th = Third quartile; IQR = Interquartile Range.
a. Multiple modes exist. The smallest value is shown

E

Distribution of Ratings: nativeness

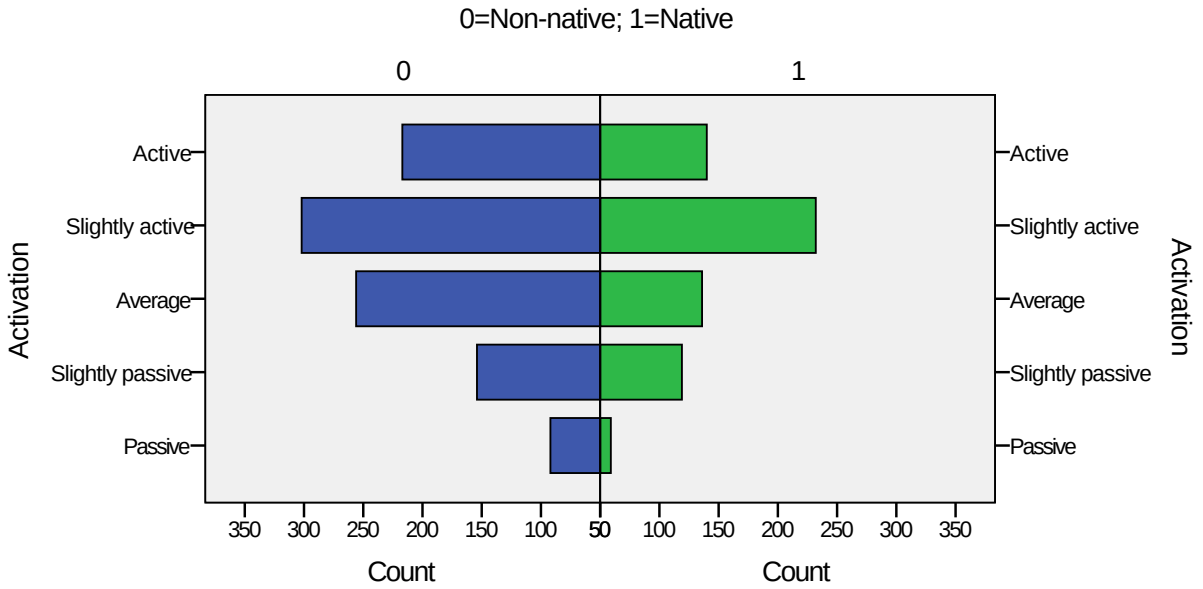


Figure E.1: Distribution of ratings on Activation scale.

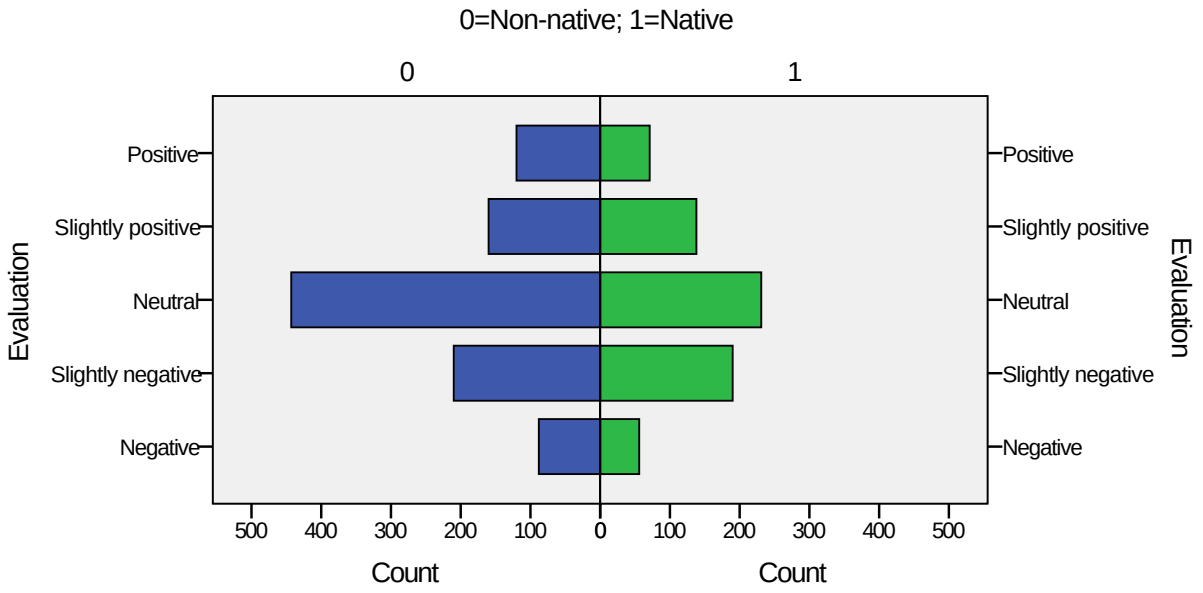


Figure E.2: Distribution of ratings on Evaluation scale.

F

Summary of Standard Deviations

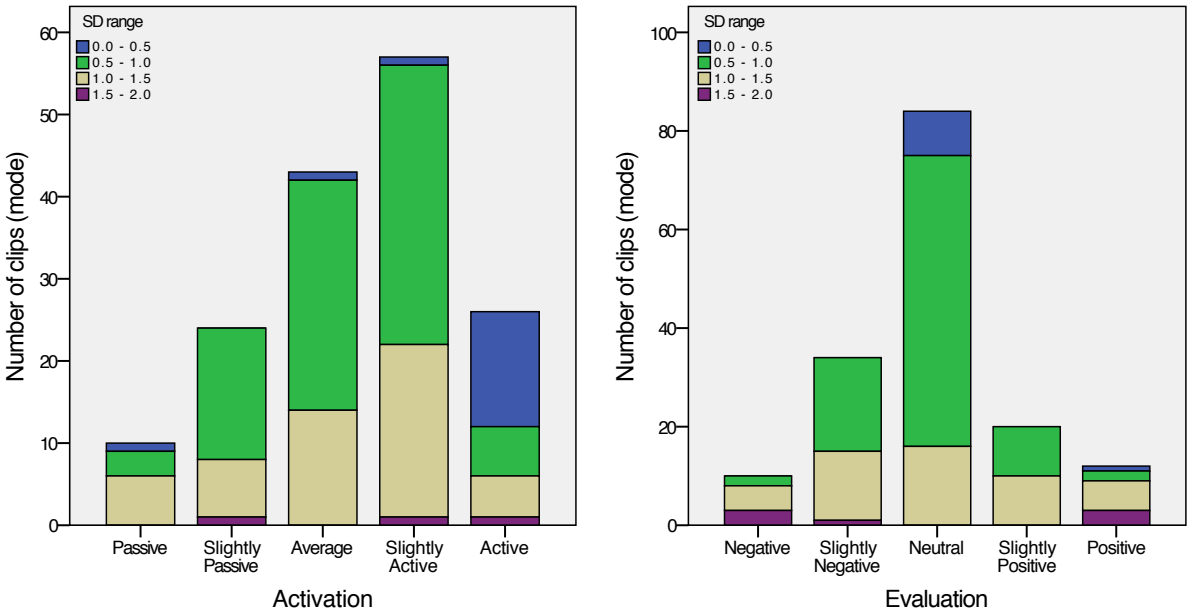


Figure F.1: Distribution of clips with respect to the SD value and the mode value obtained—the Activation (left) and Evaluation (right).

MEDIAN VALUE

	Activation					Evaluation				
SD range	Passive	Slightly Passive	Average	Slightly Active	Active	Negative	Slightly Negative	Neutral	Slightly Positive	Positive
0 to 0.5	1	0	1	1	14	0	0	9	0	1
0.5 to 1.00	2	13	30	36	6	0	17	64	9	2
1.00 to 1.5	2	6	27	15	3	1	12	24	11	3
1.5 to 2.00	0	0	1	2	0	1	0	4	0	2

Table F.1: Table shows the number of speech clips in each class—determined by the clips median values—with their respective standard deviation range.

MODE VALUE

	Activation					Evaluation				
SD range	Passive	Slightly Passive	Average	Slightly Active	Active	Negative	Slightly Negative	Neutral	Slightly Positive	Positive
0 to 0.5	1	0	1	1	14	0	0	9	0	1
0.5 to 1.00	3	16	28	34	6	2	19	59	10	2
1.00 to 1.5	6	7	14	21	5	5	14	16	10	6
1.5 to 2.00	0	1	0	1	1	3	1	0	0	3

Table F.2: Table shows the number of speech clips in each class—determined by the clips mode values—with their respective standard deviation range.

Stimuli Selection

Asset	Ratings	Margin of error	Eval_MarginofError	Act_MarginofError	Eval_GT
A1_DL_230309	15	0	0.989	0	1.968643
A8_JR_240309	9	0	1.427	0	1.156083
A3_JR_240309	7	0	1.748	0	2.945285
B6_PJ_090609	13	0.167600927	0.168	0.526	2
A6_DL_230309	14	0.209669061	0.785	0.21	3
A9_JR_240309	10	0.226215883	1.218	0.226	1.314719
A10_DL_230309	14	0.253199106	0.253	0.16	4
A9_DL_230309	15	0.253486854	0.774	0.253	3
A7_JR_240309	9	0.256222425	0.941	0.256	0.736376
A4_JR_240309	9	0.256222425	1.427	0.256	2.405666
A4_LC_260309	12	0.257019746	0.795	0.257	4
A10_PJ_090609	11	0.271760233	0.272	0.558	2.09768
A9_PJ_090609	11	0.271760233	0.628	0.272	1
A5_DL_230309	11	0.28325959	0.792	0.283	2.706804
B8_CC_200309	14	0.287100715	0.287	0.536	1.509502
B9_PA_090609	16	0.289827048	0.29	0.435	2
B7_JR_240309	14	0.301528356	0.302	0.541	1.479351
B6_EJ_250509	14	0.308623814	0.309	0.506	2
B1_JK_210809	12	0.312834431	0.633	0.313	2.719276
A3_DL_230309	12	0.312834431	1.117	0.313	1.697227
B3_PA_090609	11	0.313801377	0.314	0.69	2
B8_PA_090609	11	0.313801377	0.314	0.339	2.246938
A2_JR_240309	12	0.327170278	0.943	0.327	2
B4_JK_210809	16	0.328889591	0.329	0.597	1
B8_JR_240309	11	0.335904569	0.336	0.525	2
A5_PJ_090609	15	0.338498966	0.338	0.54	1.867226
A10_CC_200309	9	0.338951092	0.512	0.339	2
A10_JR_240309	10	0.345550291	0.92	0.346	0.762439
B10_EJ_250509	11	0.346920895	0.347	0.708	2
B9_EJ_250509	7	0.349558398	0.35	0.88	2
B1_LC_260309	14	0.352925242	0.353	0.619	2
B5_PA_090609	9	0.355824124	0.712	0.356	3.547943
B3_PJ_090609	17	0.360877507	0.361	0.473	2
B3_DL_230309	11	0.362346977	0.362	0.864	2
A7_CC_200309	14	0.375555184	0.376	0.486	1.626442
A2_DL_230309	14	0.375555184	0.751	0.376	2.7628

SELECTED BASED ON CONFIDENCE INTERVAL (CI)

H

Preliminary Survey for Filter Condition

Please listen to the speech clips as instructed by the researcher. These speech clips have been altered to make them either fully or partially incomprehensible. The task itself does not test your performance; instead, it investigates speech comprehension under certain conditions.

You will be asked to listen to 18 speech clips, if there are any parts in the speech that you can or can not comprehend, please indicate accordingly. The researcher will inform you when a new speech clip will be played to you.

For each speech clip, please answer the following question:

Can you understand any words played back to you?

Speech clip1: No ☐ Yes ☐

If yes, please transcribe:

Speech clip2: No ☐ Yes ☐

If yes, please transcribe:

Survey 1: Testing for filter condition

File ID	1	2	3	4	5	6	7	8	9	10
F_A10_CC_200309	N	N	N	N	N11	N	N	N	N	N
F_A10_DL_230309	N	N	N	N	N	N	N	N	N	Y18
F_A10_PJ_0906099	N	N	N	N	N12	N	N20	N27	N36	N
F_B10_EJ_250509	N	N	N5	N	N	N	N	N28	N37	Y19
F_B1_JK_210809	N	N	N	N	N	N	N	N	N	Y20
F_B1_LC_260309	N	N	N	N	N	N16	N	N29	N38	N
F_A1_DL_230309	N1	N	N6	Y5	N	N	N	N	N	N
F_A3_JR_240309	N	N2	N7	N8	Y9	Y13	N21	N30	N39	Y21
F_A2_JR_240309	N	N	N	N	N	N	N	N	N	N
F_A4_LC_260309	N	N	N	Y6	Y10	Y14	N22	N31	Y17	Y22
F_A8_JR_240309	N	N	Y3	N	N	N17	N23	N	N40	N
F_A10_JR_240309	Y1	N	N	Y7	Y11	Y15	Y16	N	N	N
F_A2_DL_230309	N	N3	N	N9	N13	N18	N24	N	N41	N46
F_A3_DL_230309	Y2	N	Y4	Y8	Y12	N19	N25	N32	N42	N47
F_A4_JR_240309	N	N4	N	N	N14	N	N26	N33	N43	N48
F_B3_DL_230309	N	N	N	N10	N15	N	N	N	N44	N
F_B3_PA_090609	N	N	N	N	N	N	N	N34	N	N
F_B3_PJ_090609	N	N	N	N	N	N	N	N35	N45	N

Octave above F0 min

Octave above F0 median

Octave above F0 max

N = No comprehension
Y = Some comprehension

Note: number beside indicates that user gave feedback (see next page)

Survey 1: Testing for filter condition

Notes on free-response feedback

N1 - Could make out Dublin accent
N2 - Laughter
N3 - Crying
N4 - Laughter
N5 - Gender: F
N6 - Can hear inflection
N7 - Sounded like children
N8 - Laughing
N9 - Laughing at end
N10 - Change in tone
N11 - doesn't sound like speech
N12 - doesn't sound like speech
N13 - laughing or coughing
N14 - laughing
N15 - doesn't sound like speech
N16 - sounds like singing
N17 - definitely excited
N18 - laughing
N19 - Panic
N20 - doesn't sound like speech
N21 - Crying
N22 - excitement
N23 - panic
N24 - coughing
N25 - but sounds familiar
N26 - Laughing
N27 - female
N28 - female
N29 - laugh female
N30 - female voice
N31 - female voice
N32 - female
N33 - sounds like another language
N34 - male voice
N35 - Male
N36 - music
N37 - beats
N38 - sounds like singing
N39 - shocked
N40 - Anxious
N41 - laugh
N42 - male
N43 - laughing, happy, girl
N44 - male
N45 - male
N46 - sounds like laughter
N47 - sounds upset
N48 - sounds like laughter
Y1 - "No you didn't"
Y2 - "another one"
Y3 - "...often towel"
Y4 - "another one"
Y5 - ".. Really?"
Y6 - "...it is horrendous"
Y7 - "No you didn't"
Y8 - "This is..."
Y9 - "Oh god" (or something)
Y10 - " You didn't see any of it"
Y11 - "No you didn't"
Y12 - ".... Another one"
Y13 - "...Oh no!"
Y14 - "That's nice"
Y15 - "No you didn't"
Y16 - "No you didn't"
Y17 - "I live in Ranelagh"
Y18 -"mm mm"
Y19 - "mm mm"
Y20 - "mm mm"
Y21 - "Oh god" comment: sounds upset
Y22 - "tremendous" - not sure

Survey 2: Comprehension test

		PARTICIPANT										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Result**
File ID	1	N	Y	N	Y	N	N	Y	N	N	Y	incomprehensible
	2	N	N	N	N	N	N	N	N	N	N	incomprehensible
	3	N	N	N	Y	N	N	N	N	Y	N	incomprehensible
	4	N	Y	N	Y	N	N	N	N	N	N	**comprehensible**
	5	N	N	N	N	Y	N	N	N	N	Y	incomprehensible
	6	N	N	N	N	N	N	N	N	Y	N	incomprehensible
	7	N	N	N	Y	N	N	Y	N	N	Y	incomprehensible
	8	N	N	N	N	N	N	N	N	N	N	incomprehensible
	9	N	N	N	N	N	N	N	N	N	N	incomprehensible
	10	N	N	N	N	N	N	N	N	N	N	incomprehensible
	11	N	N	N	N	N	N	Y	N	N	N	incomprehensible
	12	N	N	N	N	N	N	N	N	Y	N	incomprehensible
	13	N	N	N	Y	N	N	N	N	N	N	**comprehensible**
	14	N	N	N	N	N	Y	N	N	Y	N	incomprehensible
	15	N	N	N	Y	N	N	N	N	N	N	incomprehensible
	16	N	N	N	N	N	N	N	N	N	N	incomprehensible
	17	N	N	N	Y	N	N	N	N	N	N	incomprehensible
	18	N	N	N	Y	N	N	N	N	N	N	incomprehensible
	19	Y	Y	N	Y	N	N	Y	N	Y	Y	**comprehensible**
	20	N	N	N	N	N	N	N	N	N	N	incomprehensible
	21	N	N	N	N	N	N	Y	N	N	N	incomprehensible
	2	N	Y	N	N	N	N	N	N	N	N	incomprehensible
	23	N	N	N	N	N	N	N	N	N	N	incomprehensible
	24	N	Y	N	N	N	N	N	N	N	N	incomprehensible
	25	N	N	N	N	N	N	N	N	N	N	incomprehensible
	26	N	N	N	N	N	N	N	N	N	N	incomprehensible
	27	N	N	N	N	N	N	N	N	N	N	incomprehensible
	28	N	N	N	N	N	N	N	N	N	N	incomprehensible
	29	N	N	N	N	N	N	N	N	N	N	incomprehensible
	30	N	Y	N	Y	N	N	N	N	N	N	incomprehensible
	31	N	N	N	N	N	N	Y	N	N	N	incomprehensible
	32	N	N	N	Y	N	N	Y	N	N	N	incomprehensible
	33	N	N	N	Y	N	N	Y	N	N	Y	incomprehensible
	34	N	N	N	N	N	N	N	N	N	N	incomprehensible
	35	N	N	N	N	Y	N	Y	N	N	N	incomprehensible
	36	N	N	Y	N	N	N	N	N	N	N	incomprehensible

Survey 2: Comprehension test

File ID		Notes on results
	1	**incomprehensible** participant heard 1 out of 9 words - "you"
	2	**incomprehensible**
	3	**incomprehensible** (note: word guessed wrong &one participant almost guessed "life")
	4	2 participants judged 3 out of 6 words correct >> **comprehensible**
	5	**incomprehensible** (note: word guessed wrong)
	6	**incomprehensible**
	7	**incomprehensible** (note: word guessed wrong)
	8	**incomprehensible**
	9	**incomprehensible**
	10	**incomprehensible**
	11	**incomprehensible**
	12	**incomprehensible**
	13	1 participant heard 4 out of 13 words >> context of evaluation may be **comprehensible**
	14	**incomprehensible** (note: 1 participant guessed 1 word wrong)
	15	**incomprehensible** (note: 1 participant had a close guess)
	16	**incomprehensible**
	17	**incomprehensible** (note: 1 participant guessed 1 word wrong)
	18	**incomprehensible**
	19	3 participants judged correct >> evaluation context **comprehensible**
	20	**incomprehensible**
	21	**incomprehensible**
	2	**incomprehensible** (note: 1 participant guessed 1 word wrong)
	23	**incomprehensible**
	24	**incomprehensible** (note: 1 participant guessed all words wrong)
	25	**incomprehensible**
	26	**incomprehensible**
	27	**incomprehensible**
	28	**incomprehensible**
	29	**incomprehensible**
	30	**incomprehensible** (2 participant guessed all words wrong but **the same**)
	31	**incomprehensible**
	32	**incomprehensible** (note: 1 participant guessed 1 word wrong)
	33	**incomprehensible** (note: 2 participant guessed all words wrong)
	34	**incomprehensible**
	35	**incomprehensible** (note: 1 participant guessed all words wrong)
	36	**incomprehensible** (note: 1 participant guessed all words wrong)

Summary of Clip Parameters

FILE	F0 min	F0 mean	F0 max	F0 SD	key	range	floor	ceiling	oct_key	oct_min	oct_mean	oct_max
A10_CC_200309	160	179	234	9.9	176	73	130	350	353	321	357	467
A10_DL_230309	143	197	250	27	206	107	130	420	413	286	394	501
A10_PJ_090609	80	103	137	8.4	102	57	80	210	205	161	205	274
A1_DL_230309	120	224	443	79.8	196	323	120	460	392	240	448	885
A2_DL_230309	123	187	331	46.4	174	208	120	340	348	247	374	663
A2_JR_240309	255	357	491	63.4	349	236	230	740	698	510	713	983
A3_JR_240309	313	559	853	127.4	581	540	260	810	1162	625	1118	1706
A4_JR_240309	331	514	669	95.9	497	338	260	890	994	661	1028	1338
A4_LC_260309	180	335	486	84.5	329	306	180	720	657	360	670	971
A5_DL_230309	85	129	177	22.7	124	92	80	270	249	170	258	354
A5_PJ_090609	94	109	133	7.2	108	39	80	210	215	189	218	266
A6_DL_230309	136	173	219	15.6	171	83	130	340	342	272	346	438
A7_CC_200309	157	203	308	33.2	192	151	140	390	384	315	406	616
A7_JR_240309	221	382	529	102	358	309	220	920	716	441	764	1058
A9_DL_230309	115	155	220	22.8	151	105	110	330	302	230	311	440
A9_JR_240309	258	372	541	81.1	350	283	230	810	700	515	744	1082
A9_PJ_090609	73	130	161	23.6	137	88	70	290	275	147	260	323
B10_EJ_250509	143	192	349	27	186	206	140	420	372	287	385	698
B1_JK_210809	111	187	381	79	152	270	90	390	304	222	374	762
B1_LC_260309	164	205	347	23.8	205	184	150	410	409	327	410	695
B3_DL_230309	94	115	133	10.8	116	38	80	240	231	188	231	265
B3_PA_090609	144	166	212	13	163	68	120	330	326	288	332	424
B3_PJ_090609	89	101	126	9.3	98	37	70	210	197	178	203	252
B4_JK_210809	100	127	245	25.1	121	145	90	250	242	200	254	490
B5_PA_090609	165	273	671	102.8	243	506	150	650	486	331	547	1342
B6_EJ_250509	93	100	116	4.6	98	22	80	200	197	187	200	232
B6_PJ_090609	104	124	140	7.4	123	35	90	250	247	209	248	279
B7_JR_240309	175	213	283	18.2	209	108	160	420	418	349	425	565
B8_CC_200309	156	213	297	28.9	213	141	150	450	426	311	426	593
B8_JR_240309	144	196	279	15.3	198	135	140	390	396	287	392	558
B8_PA_090609	139	225	311	42.3	223	172	140	450	445	278	451	622
B9_EJ_250509	116	125	204	15.1	120	88	90	240	240	232	249	408
B9_PA_090609	145	170	229	11.8	168	84	130	330	336	290	341	457
AVERAGE	149.3	213.3	318.3	38.9	207.2	169.0	134.5	428.2	414.5	298.6	426.7	636.6

J

Low-pass Filter Praat Script

```
# script Low-pass filter files
# Author: John Snel
# email: john.snel@mydit.ie
# purpose: To low-pass filter all files in a given folder and output the resultant files to
a given folder.
#          Writes a report on the following values: f0min, f0mean, f0max, f0sd, range,
floor ceiling, octave_condition
# Notes: Filters 4 conditions: octave above min, mean, max, key

clearinfo
form calculate_register
        #indicate where your sound files and TextGrid are
        sentence input_folder /Users/johnsnel/Desktop/College/*Filter_CI/A/
        #indicate where you want your output to be saved for filter min_octave
        sentence output_folder_min
/Users/johnsnel/Desktop/College/*Filter_CI/A/Filter_min
        #indicate where you want your output to be saved for filter mean_octave
        sentence output_folder_mean
/Users/johnsnel/Desktop/College/*Filter_CI/A/Filter_mean
        #indicate where you want your output to be saved for filter max_octave
        sentence output_folder_max
/Users/johnsnel/Desktop/College/*Filter_CI/A/Filter_max
        #indicate where you want your output to be saved for filter key_octave
        sentence output_folder_key
/Users/johnsnel/Desktop/College/*Filter_CI/A/Filter_key

        #filter settings
        comment Filter above this frequency
        positive Smoothing 20
        comment Scale intensity to
        positive intensity 60.0

endform

# Make a listing of all the sound files in a directory.
myList = Create Strings as file list... list 'input_folder$'/*.mp3
ns = Get number of strings

# Create directories
createDirectory (output_folder_min$)
createDirectory (output_folder_mean$)
createDirectory (output_folder_max$)
createDirectory (output_folder_key$)

# Output headers to file:
# Note floor and ceiling = min_f0 and max_f0
line$="FILE'tab$"tab$'f0min'tab$'f0mean'tab$'f0max'tab$'f0sd'tab$'key'tab$'range'tab
$'floor'tab$'ceiling  octave_min'newline$'"
line$>'output_folder_min$'/infoA_min.txt
```

```
line$="FILE'tab$"'tab$'f0min'tab$'f0mean'tab$'f0max'tab$'f0sd'tab$'key'tab$'range'tab
$'floor'tab$'ceiling  octave_mean'newline$'"
line$>'output_folder_mean$'/infoA_mean.txt
line$="FILE'tab$"'tab$'f0min'tab$'f0mean'tab$'f0max'tab$'f0sd'tab$'key'tab$'range'tab
$'floor'tab$'ceiling  octave_max'newline$'"
line$>'output_folder_max$'/infoA_max.txt
line$="FILE'tab$"'tab$'f0min'tab$'f0mean'tab$'f0max'tab$'f0sd'tab$'key'tab$'range'tab
$'floor'tab$'ceiling  octave_key'newline$'"
line$>'output_folder_key$'/infoA_key.txt


for i from 1 to ns
	select Strings list
	name$ = Get string... 'i'
	Read from file... 'input_folder$'/'name$'
	mySound=selected("Sound")
	mySound$=selected$("Sound")

	pitch_step = 0.01
	To Pitch... 'pitch_step' 60 600
	myPitch=selected("Pitch")
	myPitch$=selected$("Pitch")
	minimum_f0= Get minimum... 0 0 Hertz Parabolic
	maximum_f0= Get maximum... 0 0 Hertz Parabolic
	q65 = Get quantile... 0.0 0.0 0.65 Hertz
	q15 = Get quantile... 0.0 0.0 0.15 Hertz

	max_f0 = 10*ceiling((1.92*q65)/10)
	min_f0 = 10*floor((0.83*q15)/10)

	select mySound
	# To pitch better for voice research
	;To Pitch... 'pitch_step' 'min_f0' 'max_f0'
	# for intonation research:
	To Pitch (ac)... 0 'min_f0' 15 no 0.03 0.45 0.01 0.35 0.14 'max_f0'
	myPitch2=selected("Pitch")
	myPitch2$=selected$("Pitch")

	min= Get minimum... 0 0 Hertz Parabolic
	max= Get maximum... 0 0 Hertz Parabolic
	mean=Get mean... 0 0 Hertz
	key = Get quantile... 0 0 0.5 Hertz
	span = log2(max/min)
	range = max-min
	sd= Get standard deviation... 0 0 hertz

	# Get octave values
	octave_min = min*2
	octave_mean = mean*2
	octave_max = max*2
```

```
	octave_key = key*2

	# Filter min
		# Select sound object
		select Sound 'mySound$'

		# Filter
		Filter (pass Hann band)... 0 octave_min smoothing

		# scaled files because it was clipping   ------------
		Scale peak... 0.8

		# Save resulting files
		Write to WAV file... 'output_folder_min$'/F_'name$'
		select Strings list

		# Print to txt file

	line$="'mySound$"'tab$"min:0"'tab$"mean:0"'tab$"max:0"'tab$"sd:1"'tab$"key:0'
'tab$"range:0"'tab$"min_f0:0"'tab$"max_f0:0"'tab$' 'octave_min:0"'newline$"'
		line$>>'output_folder_min$'/infoA_min.txt


	# Filter mean
		# Select sound object
		select Sound 'mySound$'

		# Filter
		Filter (pass Hann band)... 0 octave_mean smoothing

		# scaled files because it was clipping   ------------
		Scale peak... 0.8

		# Save resulting files
		Write to WAV file... 'output_folder_mean$'/F_'name$'
		select Strings list

		# Print to txt file

	line$="'mySound$"'tab$"min:0"'tab$"mean:0"'tab$"max:0"'tab$"sd:1"'tab$"key:0'
'tab$"range:0"'tab$"min_f0:0"'tab$"max_f0:0"'tab$' 'octave_mean:0"'newline$"'
		line$>>'output_folder_mean$'/infoA_mean.txt

	# Filter max
		# Select sound object
		select Sound 'mySound$'

		# Filter
		Filter (pass Hann band)... 0 octave_max smoothing
```

```
		# scaled files because it was clipping   ------------
		Scale peak... 0.8

		# Save resulting files
		Write to WAV file... 'output_folder_max$'/F_'name$'
		select Strings list

		# Print to txt file

	line$="'mySound$"'tab$"min:0"'tab$"mean:0"'tab$"max:0"'tab$"sd:1"'tab$"key:0'
'tab$"range:0"'tab$"min_f0:0"'tab$"max_f0:0"'tab$' 'octave_max:0"'newline$'"
		line$>>'output_folder_max$'/infoA_max.txt

	# Filter key
		# Select sound object
		select Sound 'mySound$'

		# Filter
		Filter (pass Hann band)... 0 octave_key smoothing

		# scaled files because it was clipping   ------------
		Scale peak... 0.8

		# Save resulting files
		Write to WAV file... 'output_folder_key$'/F_'name$'
		select Strings list

		# Print to txt file

	line$="'mySound$"'tab$"min:0"'tab$"mean:0"'tab$"max:0"'tab$"sd:1"'tab$"key:0'
'tab$"range:0"'tab$"min_f0:0"'tab$"max_f0:0"'tab$' 'octave_key:0"'newline$'"
		line$>>'output_folder_key$'/infoA_key.txt

endfor
select all
minus Strings list
 Remove
```

K

Participant Consent Forms

Participant Number: _____

Emotion Inference Listening Tests

Research Participant Release Form

You have been asked to be a participant in John Snel's ongoing speech research within the EmoVerE team, which is part of the Digital Media Centre in the Dublin Institute of Technology (DIT), Aungier Street. The entire experiment consists of two speech stimuli listening sessions. Each listening session will be performed on two occasions two weeks apart, the duration of each session being 30-45 minutes. To complete the experiment, it is necessary that you, the participant, will be able to commit to the tasks on both occasions. A dedicated time for each session will be discussed between you and the researcher and chosen most suitable for you. All data that you will be asked to provide will be kept anonymous. Your data will contribute to the overall findings of the EmoVerE project, and will be used in future academic publications. Your email address is used to provide login details for each session and monitor your ratings. The email address provided by you will be kept confidential within a password-protected database; it will not be used for any purposes other than to create an identification reference. The data provided by you will be de-identified on the second session of the experiment.

The EmoVerE research team will endeavour to do the following:

- To protect the welfare and dignity of the participant.
- To respect the individual's freedom to decline participation.
- To maintain confidentiality of research data.
- To be responsible for maintaining ethical standards.
- To take every precaution and make every effort to minimize potential risk to participants.
- To only use the data supplied by the participant with their full consent.

"I hereby give my consent to the John Snel and the EmoVerE research team in DIT to use the data gathered from my participation in this experiment for purposes of their ongoing research and future publications ***only****:"*

Name (block capitals): _____

Signature: _____

Date: _____ / _____ / _____

Research Participant Additional Information Form

Please provide the following mandatory information:

Do you have any hearing impairments: Yes ☐ No ☐

Is English First language: Yes ☐ No ☐

Please list any other languages that you are fluent or proficient in:

_____

_____

_____

The following information is **optional**, which may also be of relevance to the research being undertaken. Please feel free to include or exclude any additional information.

Age: No response: ☐

Gender: Male ☐ Female ☐ No response: ☐

Handedness: Left ☐ Right ☐ No response: ☐

L

Participant Demographics: Mixed ANOVA Analysis

To test whether two groups differed in their performance in listening to non-filtered and filtered speech (within-subjects factor), we conducted a 2 x 2 mixed ANOVA using participant nativeness, age, gender, and handedness (between-subject factor). In Table L.1, the values under $F(1, 1182)$ indicate the obtained value of the F-statistic, where 1 in (1, 1182) indicates the interaction term's degrees of freedom, and 1182 indicates the error term's degrees of freedom. The probability of obtaining the observed F-value is given by the p-value, and the effect size is indicated by η^2.

As mentioned in 8.4.1, the obtained ratings indicate violations of the assumption of normality.

	ACTIVATION				EVALUATION			
	$F(1, 1182)$	p-value	η^2	Violations	$F(1, 1182)$	p-value	η^2	Violations
Nativeness	2.181	0.14	0.001	Normality	2.181	0.14	0.001	Normality
Handedness	1.729	0.189	0.001	Normality	0.391	0.532	0.000	Normality, homogeneity of variances
Gender	0.543	0.461	0.000	Normality	0.114	0.736	0.000	Normality, homogeneity of variances
Age	0.7219	***0.007***	0.004	Normality	0.018	0.894	0.000	Normality

Table L.1: Mixed ANOVA analysis.

We can also see from the table that on the Evaluation scale, 'handedness' and 'gender' has no homogeneity of variances, as assessed by Levens Test of Homogeneity of Variance (p $>$.05). For all data there are no outliers, as assessed by inspection of a boxplot for values greater than 1.5 box-lengths from the edge of the box.

The table demonstrates that on the Activation scale, there is a statistically significant interaction between the participants *age* and the (non-filtered and filtered) conditions, $F(1,1182) = .7219$, $p = .007$, partial $\eta^2 = .004$—the effects size being very small. Although there are violations of the assumption of normality (assessed by Shapiro-Wilks test p $<$.05), there is homogeneity of variances.

M

Scatterplots for Mean and Standard Deviation

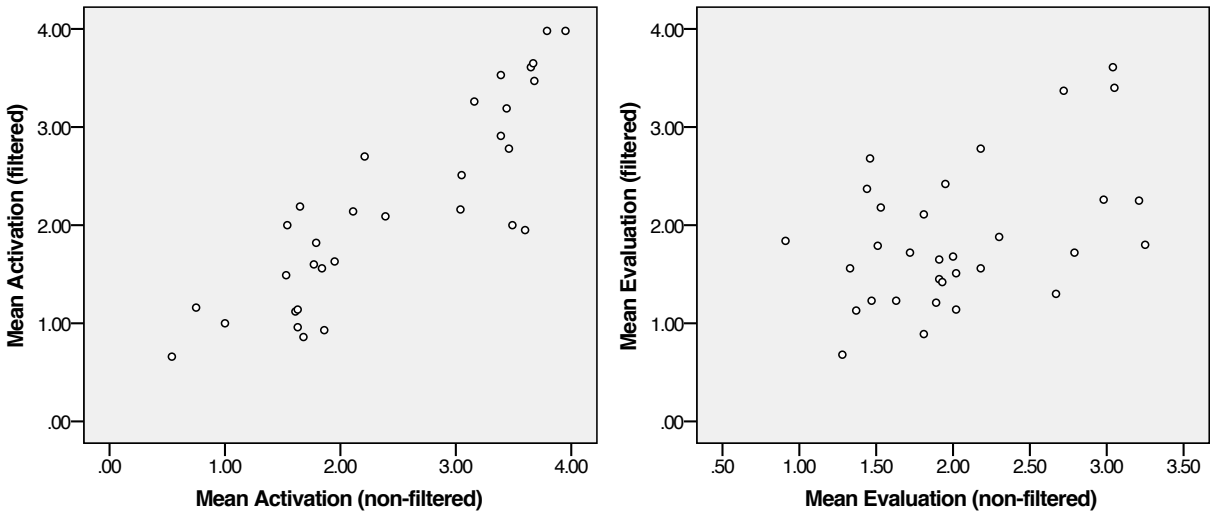


Figure M.1: Scatter Plots for Mean values.

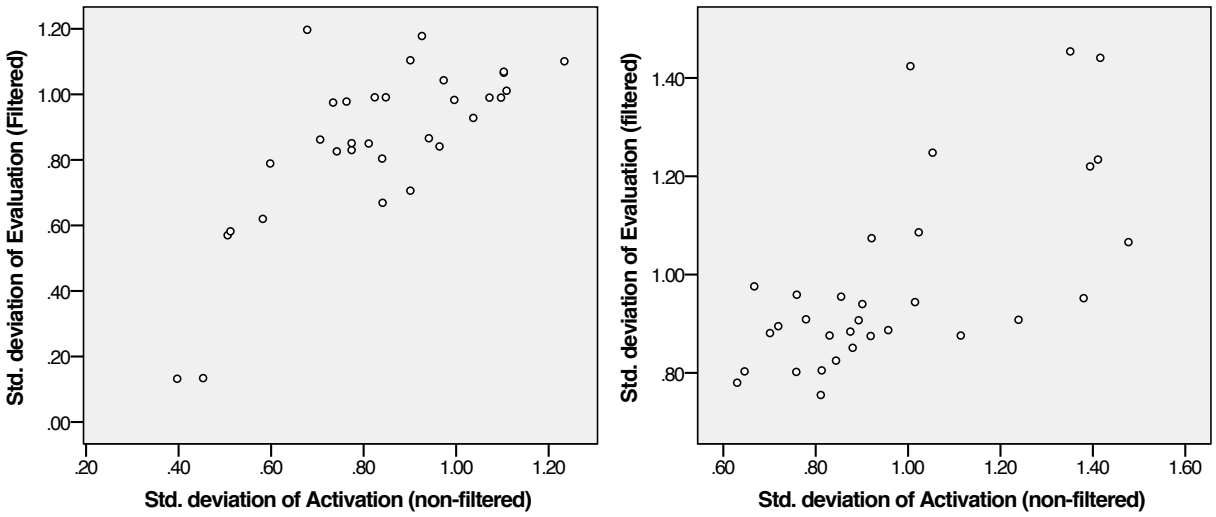


Figure M.2: Scatter Plots for standard deviation values.