2011-01-01

Naturalistic Emotional Speech Corpora with Large Scale Emotional Dimension Ratings

Brian Vaughan
Technological University Dublin, brian.vaughan@tudublin.ie

Naturalistic Emotional Speech Corpora With Large Scale Emotional Dimension Ratings

Brian Vaughan

A thesis presented to the Dublin Institute of Technology,

Digital Media Centre (DMC)

For the degree of

Doctor of Philosophy

2011

Research Supervisors:

Dr. Charlie Cullen

Prof. Eugene Coyle

Prof. Ciaran MacDonnaill

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 Institute.

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

The Institute 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 _____

Abstract

The investigation of the emotional dimensions of speech is dependent on large sets of reliable data. Existing work has been carried out on the creation of emotional speech corpora and the acoustic analysis of emotional speech and this research seeks to build upon this work while suggesting new methods and areas of potential. A review of the literature determined that a two dimensional emotional model of activation and evaluation was the ideal method for representing the emotional states expressed in speech. Two case studies were carried out to investigate methods of obtaining natural underlying emotional speech in a high quality audio environment, the results of which were used to design a final experimental procedure to elicit natural underlying emotional speech. The speech obtained in this experiment was used in the creation of a speech corpus that was underpinned by a persistent backend database that incorporated a three-tiered annotation methodology. This methodology was used to comprehensively annotate the metadata, acoustic data and emotional data of the recorded speech. Structuring the three levels of annotation and the assets in a persistent backend database allowed interactive web-based tools to be developed; a web-based listening tool was developed to obtain a large amount of ratings for the assets that were then written back to the database for analysis. Once a large amount of ratings had been obtained, statistical analysis was used to determine the dimensional rating for each asset. Acoustic analysis of the underlying emotional speech was then carried out and determined that certain acoustic parameters were correlated with the activation dimension of the dimensional model. This substantiated some of the findings in the literature review and further determined that spectral energy was strongly correlated with the activation dimension in relation to underlying emotional speech. The lack of a correlation for certain acoustic parameters in relation to the evaluation dimension was also determined, again substantiating some of the findings in the literature.

The work contained in this thesis makes a number of contributions to the field: the development of an experimental design to elicit natural underlying emotional speech in a high quality audio environment; the development and implementation of a comprehensive three-tiered corpus annotation methodology; the development and

implementation of large scale web based listening tests to rate the emotional dimensions of emotional speech; the determination that certain acoustic parameters are correlated with the activation dimension of a dimensional emotional model in relation to natural underlying emotional speech and the determination that certain acoustic parameters are not correlated with the evaluation dimension of a two-dimensional emotional model in relation to natural underlying emotional speech.

Table of Contents

Table of Figures

Table of Tables

Table of Equations

Paper List

Work contained in this PhD has contributed to a number of papers:

Cullen, C., B. Vaughan, et al. (2008a). LinguaTag: an emotional speech analysis application. Accepted paper at: The 12th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2008. Orlando, Florida, USA.

Cullen, C., B. Vaughan, et al. (2008c). Emotional Speech Corpus Construction, Annotation and Distribution. The 6th edition of the Language Resources and Evaluation Conference. Marrakech (Morocco).

Cullen, C., B. Vaughan, et al. (2008). Emotional Speech Corpora for Analysis and Media Production. 3rd International Conference on Semantic and Digital Media Technologies, SAMT. Koblenz, Germany.**:** 2.

Cullen, C., B. Vaughan, et al. (2008b). A vowel-stress emotional speech analysis method. CITSA. Genoa, Italy.

Cullen, C., Vaughan, B. ,Kousidis, S., Wang, Yi ., McDonnell, C. and Campbell, D. (2006). Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction International Conference on Multidisciplinary Information Sciences and Technologies Extremadura, Merida.

Cullen, C., Vaughan, B., Mc Auley, J., & Mc Carthy, E. (2009). "CorpVis: An Online Emotional Speech Corpora Visualisation Interface." Semantic Multimedia 5887: 169-172.

Vaughan, B., C. Cullen, et al. (2006). The Use of Task Based Mood-Induction Procedures to Generate High Quality Emotional Assets. Information technology and Communications, IT&T, Carlow, Ireland.

Vaughan, B., Kosidis, S., Cullen, C., Wang, Yi. (2007). Task-Based Mood Induction Procedures for the Elicitation of Natural Emotional Responses. The 4th International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA 2007, Orlando, Florida.

Acknowledgments

This research has taken a lot of effort and work to complete and would not have been possible without the assistance of a number of people. First and foremost I would like to sincerely thank my principal supervisor, Dr. Charlie Cullen for his incredible support and belief in me throughout the years. His patience, diligence, foresight and efforts steered me in the right direction at all times; this research would never have been undertaken or completed without him. I can only hope that this document justifies his efforts. I would also like to thank Prof. Ciaran Mc Donaill for his support and advice during this research and Prof. Eugene Coyle for his support and watchful eye over the years. I would like to thank Mr. Charlie Pritchard for his help, support and understanding throughout and particularly for his efforts in the last months of this research. I would like to thank Mr. Evin Mc Carthy for his helpful design advice and Mr. John Mc Auley for his time and valuable programming help and advice. I would also like to sincerely thank Dr. Brid Grant, Dr. Brian O'Neil and Dr. John Donovan for their invaluable support throughout this research.

I would also like to thank all my family, especially my mother and father for their generous and understanding support over the years; I could not have done it without them and I am truly thankful for everything they have done for me. Thanks to all my friends for the fun times, advice and proof reading. Thanks to all who took part in my experiments. Thanks to the rest of the SALERO group, Mr. Dermot Campbell, Yi Wang and Spyros Kousidis.

Most of all I would like to thank my wife Roisin for her patience, support and love.

1. Introduction

1.1 Motivation Of The Thesis

Emotion is, paradoxically, a widely used yet little understood concept. Considering that it is an important component of human communication, no consensus regarding the exact definition of emotional terms or emotion states exists. This is due to the complex and multimodal nature of emotional experience, and is an intrinsic part of the emotional puzzle.

The accurate recognition and understanding of another person's emotional state is vital for creating and maintaining complex social relationships. While strong primary displays of emotion are relatively easy to distinguish from normal human behaviour, it is the underlying and subtler forms of emotion that pervade the majority of our communicative processes and social interactions. Most humans are able to recognise and respond accordingly to these subtle emotional states. The increased integration of technology in our lives, evident in the ubiquitous technologies we take for granted, means that understanding and responding to the vital emotional aspects of our communicative process is an increasingly important aspect of human-computer-interaction. In order to enable various technological devices to recognise and respond to these emotional states, we must first determine a relationship between complex emotional states and measurable physical variables. Viable, structured, natural emotional data is necessary for emotional research and applications to be developed. In doing so, more meaningful and interactive technological advances can be made, with implications for the gaming industry, through the creation of complex realistic characters; for the film industry through the enhancement of virtual actors; and for medical diagnostics and assessment through a better understanding of, and ability to detect, the complex emotional states that often accompany psychiatric illnesses.

1.2 Aims Of The Thesis

1.2.1 Thesis Statement

Determining the acoustic correlates of emotional speech presents numerous difficulties. While much work has been carried out in this area, there is no conclusive methodology or a definitive set of results upon which all commentators can agree. This is due to the complex subject, the wide ranging definitions of the term ‘emotion’ and its complex multidimensional nature. Examination of the acoustic correlates of emotional speech often relies on arguably spurious data and data collection methods that mainly use actors or broadcast sources. This is further complicated through the use of subjective terms and emotional categories that have yet to be rigorously defined. This thesis seeks to address these issues and will be defended through work answering the following six research questions:

- **RQ 1: Is a two-dimensional model adequate to capture some salient aspects of natural underlying emotional speech?**
- **RQ 2: Can certain acoustic parameters of natural underlying emotional speech be correlated with the activation dimension of a two-dimensional circumplex model?**
- **RQ 3: Can certain acoustic parameters of natural underlying emotional speech be correlated with the evaluation dimension of a two-dimensional circumplex model?**
- **RQ 4: Can a practical Mood Induction Procedure (MIP) based experiment be designed and used to elicit natural underlying emotional speech from participants in a high quality audio environment?**
- **RQ 5: What are the practical considerations of annotating an emotional speech corpus?**
- **RQ 6: What are the advantages and limitations of using a large population size in rating the emotional dimensions of speech assets?**

These questions seek to investigate current methodologies via the design and implementation of a speech corpus to investigate the acoustic correlates of natural underlying emotional speech. Methods of obtaining natural emotional speech are considered along with methods of annotating said speech. A method of carrying out large scale listening tests is also considered in order to better inform the acoustic analysis of underlying emotional speech without recourse to subjective emotional terms.

Exploration of these research questions produced several original contributions to the field of emotional speech research:

1. **An experimental MIP design to elicit natural underlying emotional responses.**

2. **The use of a HD audio environment to capture natural underlying emotional speech.**

3. **The development of a three-tiered corpus annotation methodology that was used to comprehensively annotate and structure a corpus in a logical and coherent manner allowing for ease of access and utilisation via a number of web-based technologies.**

4. **The development of a large scale listening test methodology to obtain a large amount of emotional dimensional ratings for corpus assets.**

5. **The investigation of the acoustic parameters of underlying emotional states, determining that certain acoustic parameters are correlated with the activation dimension of a two-dimensional circumplex model.**

These contributions will form the basis of future investigation into emotional speech, and as such represent novel work in the field.

1.3 Contents Of The Thesis

The work of this thesis is contained within the following chapters:

Chapter 2 – Defining Emotion. This chapter reviews the literature with regard to the numerous definitions and perspectives on emotion as well as the meaning of the term compared to other terms that are often used synonymously. The biological aspects of emotion are examined to better inform methods of emotional representation. Dimensional models of emotion are considered as a method of capturing aspects of emotion experience as they are related to the activation and evaluation aspects of emotional experience without recourse to subjective emotional terms.

Chapter 3- Acoustic Parameters of Emotional Speech**.** This chapter examines certain acoustic parameters that have been found to be related to the expression of emotion in speech. Three definitions of prosody are examined to more rigidly define the acoustic approach to speech analysis. A review of the literature in relation to the acoustic parameters of emotional speech is carried out, demonstrating some consensus among researchers regarding a number of acoustic parameters in relation to certain emotional categories.

Chapter 4- Existing Emotional Speech Corpora. This chapter examines existing emotional speech corpora, arguing that the methods currently used to obtain emotional speech arguably invalidate the naturalness of the emotional speech being collected. Mood Induction Procedures are investigated as a means of obtaining natural emotional speech. Audio formats are investigated, arguing for a high quality sample and bit rate for archiving and future proofing speech recordings, with lower quality formats being used only when necessary for analysis software and listening tests.

Chapter 5- Structuring and Annotating an Emotional Speech Corpus**.** This chapter examines the annotation of emotional speech assets in a corpus, arguing for a three-tiered approach to fully annotate the metadata, acoustic data and emotional data. The IMDI schema is examined in relation to the annotation of

asset metadata. PRAAT audio analysis software and the LinguaTag application are discussed in relation to the annotation of acoustic parameters. Large-scale listening tests are examined in relation to the annotation of the emotional dimensions of the speech assets.

Chapter 6- Implementing A Three-Tiered Approach To Corpus Annotation**.** This chapter discusses the implementation and structuring of the three-tiered approach espoused in chapter 5. A persistent back end database is discussed as a method of bringing together the three-tiered annotation of the emotional assets obtained from the gaming MIP in chapter 7 and the shipwreck MIP in chapter 8. Details of the processing of the assets and contemporary World Wide Web social networks are examined as a method of obtaining emotional dimensional ratings. The development of a web-based emotional rating tool to carry out the large-scale listening tests is discussed.

Chapter 7-Case Studies: Developing An MIP To Elicit Natural Emotional Speech**.** This chapter details the development of an MIP to elicit natural emotional speech in a high quality audio environment. Two case studies were carried out to test the recording equipment and experimental setup. Ratings were obtained for the assets from the gaming MIP. The results of the case study and the ratings received were used to create a new final MIP in chapter 8 along with a revised method for obtaining ratings.

Chapter 8- A Final Task-Based MIP To Elicit Naturalistic Underlying Emotional Speech**.** This chapter examines a final MIP designed to elicit natural underlying emotional speech and a method of obtaining emotional ratings using contemporary web based social networks. The ratings received were subjected to a statistical analysis procedure to determine their dimensional rating and position on each dimension of the circumplex model used. Recommendations for the use of MIPs and large scale listening tests are discussed.

Chapter 9- Analysis Of The Acoustic Parameters Of Natural Underlying Emotional Speech Assets**.** This chapter examines the acoustic analysis of a set of assets taken from the shipwreck MIP experiment in chapter 8. A correlation was

found for certain acoustic parameters and the activation dimension of the circumplex model. The results of the chapter are discussed in relation to the findings of chapter 3.

Chapter 10- Conclusions**.** This chapter concludes the research and summarises all the work undertaken within the thesis, detailing the reasons why each aspect of the research was carried out and discusses the research questions that were asked throughout the thesis. Future work in a number of areas related to the work carried out is considered.

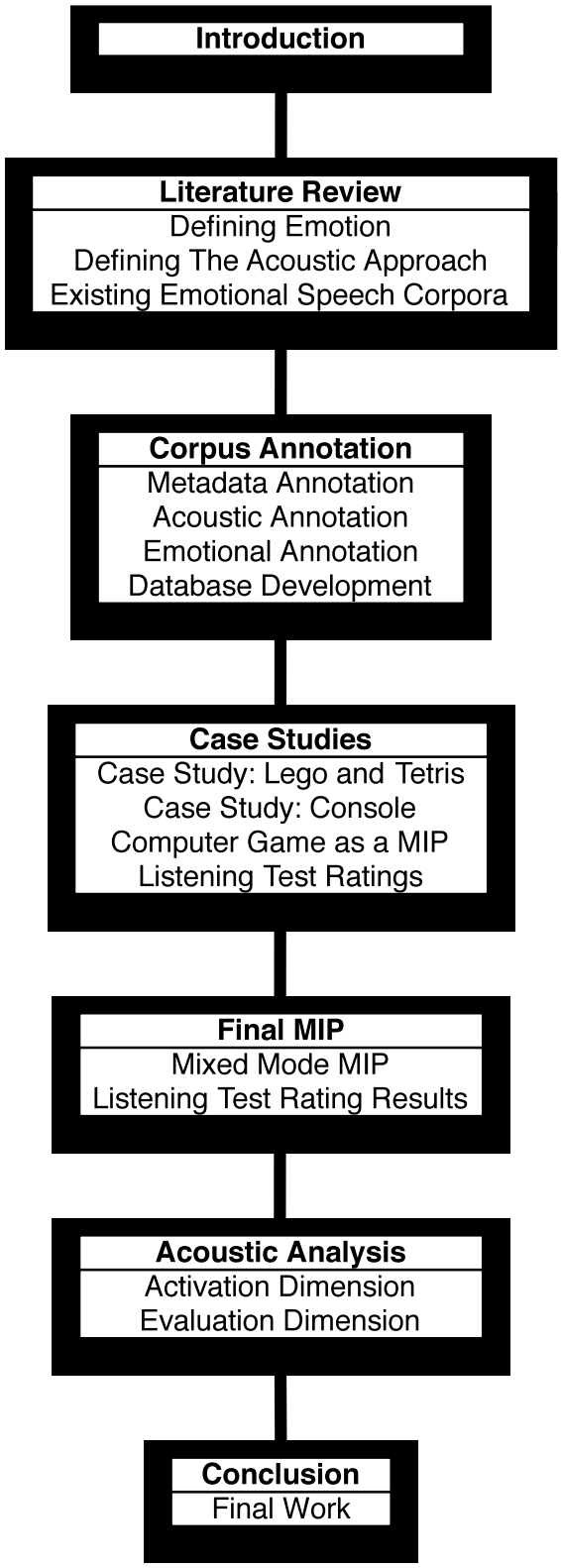


Figure 1: Flow diagram illustrating the structure of this thesis. The literature review discusses the theoretical background of emotion, examines the acoustic parameters of emotional speech and considers existing speech corpora along with the methods used to create them. The literature review is used to formulate six research questions that are then answered in subsequent chapters.

2. Defining Emotion

The purpose of this chapter is to examine current theories and perspectives on emotion. Emotion is a multi-component entity, arguably having physiological and psychological attributes as well as behavioural, social and cognitive elements. This chapter seeks to define emotion through examination of the different approaches and definitions as well as examining methods for emotional representation. One of the biggest problems in the study of any aspect of emotion is the use of subjective terminology to describe emotional states; there is no concise definition of fear or anger for example, yet these terms are frequently used throughout the literature. There is no objective method for determining if a subjective emotional term has the same meaning for different researchers. Humans use a myriad of terms to describe emotional states: angry, annoyed, irritated etc. Are these terms synonymous or descriptions of distinct emotional states? Therein lies one of the fundamental problems in the study of emotion. While some commentators would argue that there are as many emotions as there are descriptions for them, often resulting in long lists of descriptive terms, others argue that there are a few basic emotions that underpin all other emotions.

In this chapter the four main psychological perspectives on emotion are first considered (2.1), followed by an examination on the biological basis of emotion (2.2). Full-blown and primary emotions as well as their antecedents are next considered, arguing that underlying emotional states are constituent parts of the communicative process and thus warrant investigation (2.3 and 2.4). Dimensional emotional models are then examined, and it is argued that they are an advantageous method of avoiding the use of subjective terminology in describing emotional states, while also making recourse to the physiological and cognitive aspects of emotion (2.5.3). Finally, the differentiation between the terms emotion, mood and affect is discussed (2.6).

2.1 The Four Main Emotional Perspectives

This section considers the four emotional perspectives prevalent in contemporary psychology: the Darwinian, Jamesian, Cognitive and Social Constructivist

perspectives (Cornelius 1996). Each of the four perspectives postulates about the nature, origin and purpose of emotion, in relation to human interaction and society.

2.1.1 The Darwinian Perspective

The Darwinian perspective contends that emotions are reaction patterns which have become shaped by evolutionary experiences, allowing for the survival of certain environmental dangers (Darwin 1872). Within this perspective emotions are common to the entire human species, serving a basic survival function and eliciting an evolved response in order to best deal with a given situation (Plutchik 1980). Therefore the same emotions should be observable in all humans, with research showing evidence for the universality of a small number of facial expressions; happiness, sadness, fear, disgust, anger and surprise (Ekman 1987). These facial expressions can be considered basic/primary emotions that represent basic evolved survival based response patterns, with all other emotions arguably being derived from them (Cornelius, Randolph R. 2000). While Ekman's research suggests that these basic emotional facial expressions are culturally universal, some researchers disagree. Jack et al examined the decoding of facial expressions between Eastern and Western observers and found that the decoding strategy of the Eastern observers focused on the eye region, whereas Western observers focused on the face as a whole. This led to Eastern observers being more prone to ambiguous information and confusion regarding the exact emotional expression (Jack, Blais et al. 2009), suggesting that emotional facial expressions may be culturally relative. In contrast, Matsumoto and Willingham et al. examined the facial expressions of athletes from a wide range of countries at the 2004 Athens Olympic Games (Matsumoto, Willingham et al. 2009). They found that while the athlete's initial emotional displays corresponded to a set of basic universal emotions they quickly modulated into culturally influenced expressions. This is arguably a demonstration of the important social aspect of emotion (2.1.4), indicative of the possibility that emotional displays can be universal whilst also being culturally determined. Moreover, Shaver carried out a study that asked people from three different cultures (American, Chinese and Italian) to sort a list of emotions into groups based on their similarity to each other; cluster analysis of these results suggested that six emotions (love, joy, surprise, anger, sadness and fear) could be described as basic emotion categories, with a high degree of overlap between the three cultures (Shaver, Schwartz et al. 1987; Shaver, Wu et al. 1992). These six emotions

are very similar to the basic six described by Ekman (happiness, sadness, fear, disgust, anger and surprise) (Ekman 1987). While there are differences between Shaver and Ekman there is agreement between the two regarding four basic emotions: fear, anger, sadness and happiness/joy. Similarly, Scherer et al. conducted research into the universality of vocal emotion across nine countries which yielded promising results; nevertheless, German actors were used to create German sounding unintelligible speech samples, and accuracy across the nine countries decreased as the dissimilarity in languages increased (Scherer, Banse et al. 2001b).

While there is strong evidence for the universality of facial expressions, and some evidence for the universality of vocal expressions of emotion, the important aspect of the Darwinian perspective is that it firmly roots emotions in biological activation, via facial expressions or speech.

2.1.2 The Jamesian Perspective

The Jamesian perspective continues the ideas found in the Darwinian tradition, suggesting that physiological changes in the body will engender emotional states (James 1884)[1], thus also firmly grounding emotions within biological activation. In the Jamesian perspective it is the perception of the physiological changes in our bodies that we describe as emotion (Thambirajah 2004).

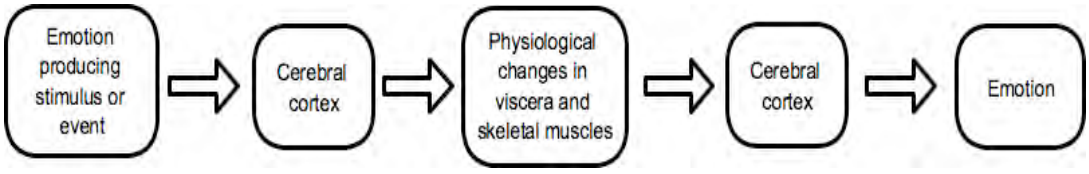


Figure 2: The James-Lange theory of emotions, adapted from (Thambirajah 2004). A stimulus or event stimulates the cerebral cortex, which in turn induces physiological changes in the internal organs. These changes are relayed to the cerebral cortex and are experienced as emotion.

Within this perspective, our bodies react automatically to events in the environment around us; we are predisposed to react in a certain way to environmental features, similar to the Darwinian notion of emotions being evolved survival patterns (Cornelius, R.R 2000). Our experience of the physiological changes that our bodies

[1] Carl Lange had a similar theory to James within the same time period and so the perpective is sometimes referred to as the James-Lange Theory.

go through constitute what we know as emotions and thus every emotion has a unique corresponding physical response; research into whether certain emotional states can be distinguished from each other using physiological measurements has been inconclusive in this regard (2.2.2). While contemporary theories on the neurological and physiological aspects of emotion suggest that there is a bodily aspect to emotion, contemporary findings suggest that they are a lot more complicated than the Jamesian perspective postulates (2.2.1 and 2.2.2).

The idea that bodily changes are experienced as emotion is the basis of the facial feedback hypothesis; assuming a facial expression of a certain type is believed to cause a change in a persons emotional state depending on the emotion connected to the facial expression (smiling will engender happiness etc) (Cappella 1993). Strack demonstrated this with a set of experiments that asked participants to rate how funny they found a series of cartoons while holding a pen in their mouths. Those that held it in their teeth, thus assuming an expression similar to that of a smile, reported finding the cartoons funnier than those that held the pen in their lips which caused a facial expression more akin to frowning (Strack 1988). Conversely, Schneider and Josephs found that children smiled more after receiving negative feedback than they did when they received positive feedback (Schneider and Josephs 1991), thus suggesting that smiling may not always be correlated with positive emotional states. However one must keep in mind that this may be due to developmental and/or social aspects. Davis, though finding evidence to support the facial feedback hypothesis, contends that there may be other factors that need to be considered such as cultural display rules that may influence or inhibit facial expression (Davis 2009). Similarly Matsumoto's findings regarding the segueing of universal displays of facial emotion into culturally relative norms would also suggest that cultural and social display rules are an important element in considering results from the facial feedback hypothesis (Matsumoto, Willingham et al. 2009).

Both the Darwinian and Jamesian perspectives ground emotions in bodily states: without the physical body emotions do not exist, as each emotion is associated with and accompanied by a level of biological activation. Both perspectives view emotions as serving important survival functions with Darwinians and Jamesians often referring to emotions as “action tendencies”; emotions induce in an organism the tendency to

act in a certain way under certain conditions (‘fight’ or ‘flight’) (Frijda 1986). The most salient aspect of the two perspectives is the emphasis on the physiological aspect of the emotional experience.

2.1.3 The Cognitive Perspective

The Cognitive perspective centres on the concept of appraisal: an evaluation of stimuli received through low-level cognitive processes (Arnold 1960). All emotions within this perspective are a consequence of these appraisals; the emotions people feel and experience are dependent on whether they judge a particular environmental situation to be, broadly speaking, good or bad (Arnold 1960; Cornelius, R.R 2000; Scherer, Johnstone et al. 2003). External stimuli in the environment cause humans to appraise the situation to determine a proper course of action, with the resulting emotions being the outcome of this cognitive appraisal. Therefore, every emotion corresponds to a unique appraisal of a situation that, if changed, will cause the corresponding emotion to also change: thus an initial emotional response to an event can be changed by reappraising the event (Spiesman 1964).

Plutchik argues that cognitive processes evolved in order to predict the future more effectively and allow for a better adaptation to the environment; his feedback loop theory postulates that a situation can be reappraised based on changes in the external environment, thus leading to a different emotional response (Plutchik 2001). Scherer developed an appraisal centric model that he termed the component process model (Scherer 1984a): this has allowed physiological predictions to be made about vocal changes due to certain emotions (Banse and Scherer 1996). The component process model consists of five different processes: cognitive, physiological, motivational, expressive and feeling. Within the cognitive component of the model, appraisal of a situation consists of a sequence of stimulus evaluation checks (SECs). The SECs are: novelty check, intrinsic pleasantness check, goal/need conduciveness check, coping potential check and norm/self compatibility check (Scherer 1984a). These SECs are a framework describing how the process of appraisal takes place within the human cognitive system; they operate in a temporal sequence with the resulting emotion being a product of the different SEC outcomes. A stimulus is appraised to see whether it is worthy of attention; if it is pleasant (resulting in pleasure or pain), whether it meets or satisfies a goal or need; the ability of the organism to adjust to the

stimulus/situation; and how significant the stimulus is to a persons self-concept and to social norms and values (whether or not it is socially acceptable) (Scherer 2001a).

The cognitive perspective puts the appraisal process at the centre of emotional states: similar to the Jamesian and Darwinian standpoint that emotions are action-tendencies, the appraisal process informs us about environmental features leading to a state of bodily activation (Cornelius, Randolph R. 2000).

2.1.4 The Social Constructivist Perspective

The Social Constructivist perspective defines emotions as socially constructed patterns that are learned and shared within cultures (Averill 1980). Social Constructivism determines emotions as a means of regulating social interaction and are constructed to serve specific *social* purposes, although biological foundations are recognised as being of secondary importance (Cornelius, R.R 2000). Rules specific to each culture determine how emotions are experienced and expressed. Emotional reactions to environmental stimuli are socially constrained and constructed, being simply responses to socially determined appraisals (Averill 1980). Most importantly, emotions have an important social function in that they are necessary for successful social interaction: they allow for the negotiation and survival of complex social situations and relationships. The Social Constructivist Perspective places the Darwinian, Jamesian and Cognitive perspectives in a social context that is culturally relative. While there is evidence that there are cross-cultural universals (2.1.1), how an emotion is expressed may be culturally determined. Work by Matsumoto et al suggests that this is the case: while there are universal emotional expressions, cultural and social display rules quickly come into effect (Matsumoto, Willingham et al. 2009). The stimulus evaluation checks in Scherer's component process model includes a norm/self check to check how a stimulus affects internal ideas of self and how socially acceptable the stimulus and resulting reaction are (Scherer 2001a). The social constructivist perspective recognises that emotions have a basic evolutionary survival related function, but more importantly they have an important social and cultural survival role.

Within this perspective the primary role of emotion is to maintain and regulate complex human social relationships. While the biological foundation of emotions is

accepted, it is the social role of emotions that is of the most importance (Schröder 2004b).

2.1.5 Comparison Of Emotional Perspectives

The four main perspectives are different approaches to the problem of how to define, explain and, ultimately, examine emotion. While they do differ in their approach there is a lot of cross-pollination between them. The Darwinian and Jamesian perspectives are very similar, with contemporary Darwinians and Jamesians referring to emotions as 'action tendencies'(Frijda 1986). The first three perspectives (Darwinian, Jamesian and Cognitive) are arguably complimentary, working together to form a more cohesive view of emotion: the mind makes an appraisal of a situation which induces a physiological state of activation in the body; this state serves a survival function in preparing the body in the best possible way to survive the situation. Thus, the criteria of all three perspectives can be satisfied: the emotion is based on both an appraisal of a situation and a bodily state (a result of the appraisal) and ultimately the emotion has a survival function. The Social Constructivist perspective can be seen to utilise and encompass the theories of the other three. Cornelius contends that the first three perspectives are on a path of convergence, being integrated into contemporary studies of emotion: any appraisal and corresponding bodily activation are conceivably constrained by cultural and societal rules (2.1.4) (Cornelius, R.R 2000). Research regarding the constraining of universal emotional displays via cultural display rules seems to support the Darwinian/Jamesian and Social Constructivist perspectives (Ekman 1987; Matsumoto, Willingham et al. 2009).

The Social Constructivist perspective encapsulates the other three within its broader perspective as *"The scale of the definition is such that it encompasses the definitions offered by the other perspectives"* (Cornelius, R.R 2000); the Darwinian, Jamesian and Cognitive perspectives conceivably operate within the wider Social Constructivist framework, providing a more cohesive perspective overall. This argument for the integration of all four of the emotional perspectives is strengthened by the fact that they can all lay some claim to basis in fact or experimental demonstration (Ekman 1993; Cornelius, R.R 2000; Matsumoto, Willingham et al. 2009). Having considered the four main perspectives on emotion in contemporary psychology, the biological basis of emotion is considered.

2.2 The Biological Basis of Emotion

The four perspectives examined in the previous section often make recourse to biological (action tendencies, physiological changes) and neurological processes (cognitive appraisal). While the Darwinian and Jamesian perspectives are the two most concerned with the physiological aspect of emotions, the cognitive perspective sees bodily activation as the result of cognitive appraisal, and the social constructivist perspective recognises the biological basis of emotions but sees their social function as being of greater importance. Although the nature of cognition in relation to neurological processes is beyond the scope of this research, appraisal of a given situation arguably involves a neurological element: appraisal can only be made using the information received through the sensory organs and processed accordingly in the brain. Likewise, physiological responses to external and internal stimuli originate in, and are controlled by, neurological structures and processes via the nervous system, specifically the Autonomic Nervous System (ANS) and the Somatic Nervous System (SNS).

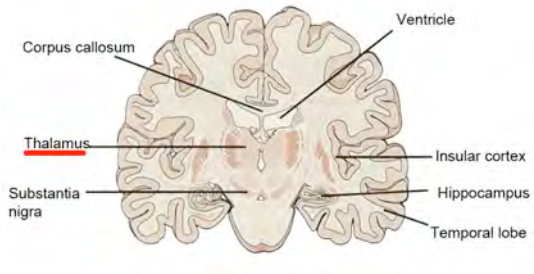
The ANS is primarily concerned with the involuntary reflexes of the internal organs (visceral reflexes) and behaviours of the body, while the SNS is primarily concerned with the movement of skeletal muscle and voluntary bodily functions. Emotion related changes of the ANS affect certain bodily faculties: heart rate, respiration, sweat glands (skin conductance), bladder contraction, salivation etc. (Kandel 2000; Fitzgerald 2002; Gazzaniga 2009). As with the neurological aspect of emotion, an exhaustive examination of physiological responses in relation to neurological processes and structures are beyond the scope of this research. However, evidence suggests that there is a physiological and neurological dimension to emotion. This has important implications in implementing methods of measuring and classifying emotional states in participants, especially models that allude to the basic physiological and neurological aspects of emotional states.

2.2.1 The Neurological Aspects Of Emotion

One of the earliest theories regarding the neurological aspects of emotions was the Cannon-Bard theory, and was originally put forward as an alternative to the James-Lange theory (Strongman, K.T. 2003). The basic theory is that the emotional process

is essentially thalamic in nature: an external situation or event stimulates bodily receptors (sight, sound, taste etc) that pass through the thalamus, which in turn relays the signals to the cerebral cortex; the cortex then stimulates thalamic processes, thus affecting the bodily receptors.

(a)



(b)

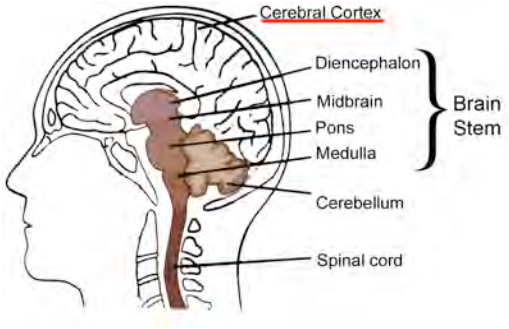


Figure 3: Diagrams showing the location of the a) thalamus in a cross section of the brain, b) the cerebral cortex in an elevation view of the head and brain, adapted from (Kandel 2000)

The cerebral cortex (Figure 3a) is the area of the brain were the majority of operations relating to our cognitive abilities occur, while the thalamus, contained in the Diencephalon structure in Figure 3a and detailed more clearly in Figure 3b, processes the majority of information from the central nervous system to the cerebral cortex (Kandel 2000). It was thought that there was a link between these thalamic processes and particular emotional displays, with the thalamic reaction creating a bodily response and simultaneous emotional experience (Puri 1998; Strongman, K, . T. 2003; Gazzaniga 2009).

James Papez also emphasised the neurophysiological aspect of emotion: he proposed that a circuit of certain neurological structures worked together to control emotional states (Papez 1937)[2]. This collection of neurological structures was later termed the limbic system by Paul McLean, and extended to include other neurological structures including parts of the hypothalamus, and the amygdala, with contemporary research showing that further structures are also involved (Figure 4) (Kandel 2000).

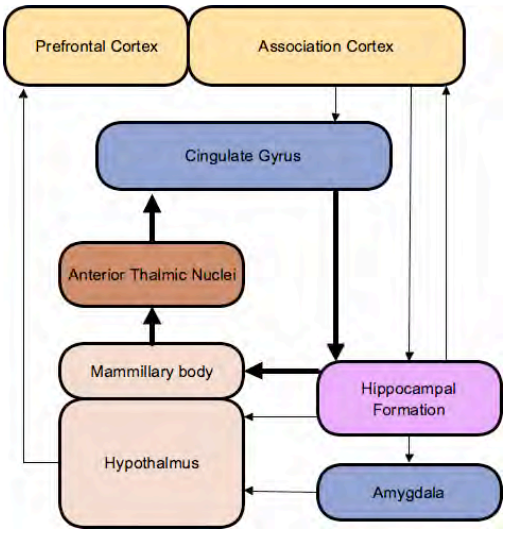
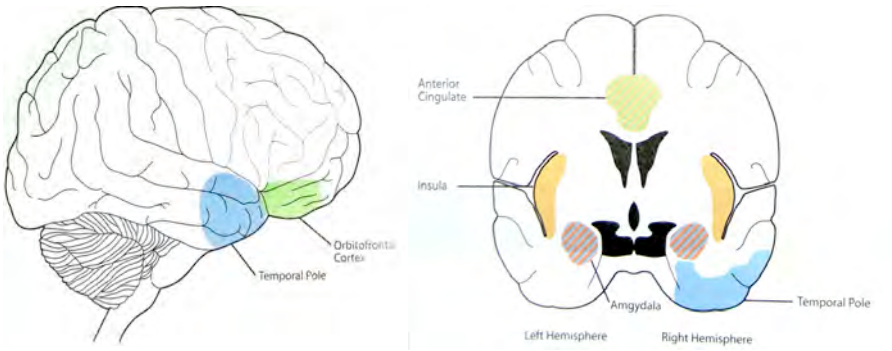


Figure 4: The Papez circuit: the original circuit is indicated by the thicker lines while the extended circuit is indicated by the thinner grey lines, adapted from Kandel (Kandel 2000).

While the limbic system is still referred to in contemporary literature, the original theories of Papez, MacLean and the Cannon-bard theory have not been fully supported by more recent research; contemporary theories suggest that there is more than one neural circuit or system involved in emotional states (Gazzaniga 2009). Research has suggested that the amygdala is closely involved with the emotional state most commonly described as fear as electrical stimulation of the amygdala induced fear and apprehension in subjects (Kandel 2000). Further research using functional magnetic resonance imaging (fMRI) has shown that the amygdala is also involved in the recognition of facial expressions as well as being directly involved in fear conditioning and learned emotional responses. Figure 5 illustrates the neurological structures that current research has found to be associated with particular emotions.

[2] Commonly referred to as the Papez circuit.



Emotion	Associated Brain Area	Functional Role
Fear	**Amygdala**	**Learning, Avoidance**
Anger	**Orbitofrontal Cortex, Anterior Cingulate Cortex**	**Indicate Social Violations**
Sadness	**Amygdala, Right Temporal Pole**	**Withdraw**
Disgust	**Anterior Insula, Anterior Cingulate Cortex**	**Avoidance**

Figure 5: Areas of the brain thought to be associated with certain emotional states. Adapted from (Gazzaniga 2009)

As discussed, fear is associated with the amygdala and a learned fear response. Anger is associated with the orbitofrontal cortex and the anterior cingulate cortex, which are believed to determine if a violation of social convention has taken place. Sadness is also associated with the amygdala along with the right temporal pole: both areas are believed to be associated with withdrawal. Disgust is associated with the anterior insula and with the anterior cingulate cortex, which are associated with avoidance. While these findings are not irrefutable, they do demonstrate that the neurological processes involved in the expression of emotion are far more complex than Cannon-Bard, Papez or Mc Lean hypothesised. However, these findings are not conclusive and are, at most, only indicative of a potential relationship between certain brain structures and emotion related behaviour.

2.2.2 The Physiological Aspects Of Emotion

In 1953, Ax examined fear and anger to determine if there were physiological differences between them (Ax 1953). To induce a state of fear, subjects were given a

mild electric shock followed by a staged equipment failure: sparks were triggered by the researcher who acted surprised and alarmed and suggested that the wiring was faulty and potentially dangerous. After a period of rest, anger was induced by a polygraph operator who was described to the participants as incompetent; the operator also made sarcastic comments and acted rude to the participants and other researchers. Music, chosen by each participant, was used to prior to the experiment and between the two forms of emotional induction to relax the participants. Ax noted that some previous research did find physiological differences between certain emotional states but they did not establish if these changes are due to the emotional state or were the distinctive physiological patterns of the individual. Ax recorded haemodynamic response (circulation), respiration, skin temperature (on the face and fingers), galvanic response (skin conductance), blood pressure and the activity of the frontalis muscle[3]. He found that anger caused a greater than average reaction in certain physiological measurements: blood pressure was increased, heart rate was decreased, muscle tension increased, and there was an increase in the galvanic skin response. For fear he found that there was a greater than average increase in galvanic response, the number of muscle tension peaks and an increase in respiration rate.

Ax's study is narrow in that it only focuses on two emotional categories but it did include 43 test subjects and some of his results were replicated in later studies. Peter & Hebron carried out a review of 13 studies regarding the physiological aspects of emotion (Peter and Herbon 2006). From this they found that heart rate (HR) and skin conductance level (SCL) were two physiological variables that demonstrated the most consistency in relation to fear. Other variables measured did not show such a strong correlation and the findings of various researchers are often contradictory: Ax's findings regarding HR in relation to anger are inconsistent with the findings of other researchers who mainly found HR to increase with anger (Fredrickson 2000; Christie 2002). Physiological findings in relation to other emotions are also relatively inconclusive (Peter and Herbon 2006). The lack of a consensus may, in part, be due to the experimental methodology or differences in the understanding of the terms anger and fear (sections 2.3 and 2.4). However, one must consider the point raised by Ax,

[3] The frontalis is a muscle of the scalp, above the eye-brows, that serves to raise or lower them accordingly, as well as cause creases in the forehead (Netter 1997; Sinnatamby 1999)

that differences in physiological measurements might be due to distinctive physiological patterns in individuals.

Oehme et al. used five films to induce a number of different emotional states, spanning the four quadrants of a three dimensional model of arousal, valence[4] and dominance[5]. They measured skin conductance level (SCL), heart rate (HR), facial electromyogram (EMG), breathing rate (BR), skin temperature and electrodermal activity (EDA) (Oehme 2007). The dimensional model was used to rate the emotional state of participants. Their results indicated that SCL, HR and BR are important in differentiating certain emotional states: SCL and BR were correlated with the arousal dimension while HR was correlated with the valence dimension of the model used. Overall the results suggest that there are physiological differences between certain contrasting emotional states but that no specific physiological patterns relate to specific emotional states. Oeheme et al's study is interesting as it correlated physiological measurements with the axes of an emotional dimensional model. Similarly, some earlier studies also correlated physiological measurements with emotional dimensions: Bradly found a strong correlation between valence, HR and zygomatic[6] EMG (ZEMG) while Detenber found SCL was correlated with arousal/activation (Bradley 1993; Detenber 1998). Anttonen found HR decreased with all stimuli (positive and negative) with the biggest decrease occurring when negative stimulations were used, this suggested that HR was correlated with valence, similar to Oheme et al's findings (Anttonen, Surakka et al. 2009).

However, Peter and Herbon found that no consideration was given to analysing individual differences in the literature they reviewed, echoing Ax's concern in determining whether physiological changes are due to an emotional state or are unique to the individuals being studied (Peter and Herbon 2006). This aspect is not usually considered in relation to the physiological correlates of emotion thus making a

[4] Valence defines how good or bad, or positive or negative something is judged to be. Dimensional models are discussed further on in this chapter (section 2.5).

[5] Dominance was used to distinguish between similarly rated emotional states and measured the strength of dominance using a dimension comprised of weak and strong at opposing ends.

[6] The muscles at the corner of the mouth (Netter 1997).

wider generalisation regarding the physiological correlates of emotional states difficult.

2.2.3 Discussion Of The Biological Basis Of Emotion

While the exact neurological processes involved in emotional expression have yet to be conclusively determined, it can be argued that there is enough research and evidence to suggest that a number of different neurological structures are involved in the expression, feeling and display of emotion. While no conclusive pattern of physiological changes has been determined for singular emotional states, but as with the neurological aspect of emotion, research has demonstrated that there is a physiological aspect to emotion. Furthermore, physiological changes brought on by emotion have been found to directly affect the organs of speech production and the speech produced (Oudeyer 2003; Breazeal 2004). The inconclusive findings of the physiological correlates of distinct emotional states could potentially be explained by Schachter's and Singer's two-factor theory (Schachter 1962). This postulates that although emotional states may have similar physiological correlates, it is our appraisal of the cause of the physiological changes that determines their emotional label, suggesting that the same physiological states can be responsible for, or a factor in, more than one emotional state (Strongman, K.T. 2003; Thambirajah 2004). Methodological differences and a lack of a concrete definition for emotional terms (anger, fear etc) may prevent any generalisation or consensus being reached. Despite the contradictions and methodological differences regarding the physiological aspects of emotions and the indefinite findings regarding emotional neurological processes and structures, it is argued that there is enough evidence to suggest that any study of emotional states should make recourse to these neurological and physiological aspects.

While the four perspectives and the biological basis of emotion are concerned with its form and function, consideration must also be given to the definition of the term 'emotion' and emotion terms used throughout the literature.

2.3 Full-blown And Underlying Emotions

Most theoretical definitions of emotion will make recourse to full-blown emotions, often as a means of indicating source or hierarchy for certain emotional states. Cowie & Cornelius argue that full-blown emotions are multi-faceted involving numerous aspects: appraisal of a situation (Arnold 1960); physiological changes (Scherer 1986b); action tendencies (Frijda 1986); subjective emotional feeling (Russell 1980) and expressive physical behaviour (Ekman 1993). Similarly, Schröder contends that emotions are complex phenomena with no one study covering all possible aspects: the four perspectives previously discussed focus on particular aspects (biological, cognitive and social) (Schröder 2004b). Fontaine and Scherer et al. argue that there are six principle components of emotion: appraisal, psychophysiological changes, motor expressions, action tendencies, subjective experiences and emotion regulation (see 2.5.2 for further discussion on this) (Fontaine, Scherer et al. 2007). Similar to Frijdas action tendencies, Sloman postulates that full-blown emotions are part of a hybrid bodily mechanism, spurring the body into action by acting as alarm signals (Frijda 1986; Sloman 1998). It has been argued that emotion is a constituent of nearly all cognitive states ***(Buck 1999; Lazarus 1999; Cowie and Cornelius 2003), with Cowie and Cornelius using the term ‘underlying emotion’ to refer to the everyday emotional states that underlie our interactions with the world around us: they describe experiments which indicated that primary emotional states interrupt speech and induce a level of incoherency, similar to Slomans interrupting ‘alarm signal’ view. Since emotions are a central part of human communication and interaction, the majority of everyday emotional experiences are arguably underlying in nature as they do not interrupt speech or induce incoherency. Although often given less theoretical consideration than full-blown emotions, it has been argued that underlying emotions are an important aspect of spoken interactions and are milder emotional states in comparison (Laukka, Juslin et al. 2005). If full-blown emotions are partially characterised by their intensity and disrupting nature, then underlying emotions are full-blown emotions.

2.4 Primary and Secondary Emotions

Similar to the idea of full-blown and underlying emotions is the notion of primary and secondary emotions: the distinction between the two sets of terms is not always

obvious and thus they are often used synonymously i.e primary/full-blown and secondary/underlying. The notion of primary and secondary emotions can be considered a 'palette theory' of emotion: secondary emotions are formed by mixing the primary emotions in various strengths (Scherer 1984b; Plutchik 2001) (see 2.5 for related discussion). Similarly, Shaver describes a prototype model of emotion in which five basic (prototype) emotional categories are proposed (anger, love, joy, fear and sadness): in this approach, these basic emotions are considered prototypes for other emotions: an emotion such as joy would be the prototype for contentment, pride and zest. (Shaver, Schwartz et al. 1987) Similar to the palette theory, prototype emotions are blended to form various other emotions. This theory is in agreement with the cognitive perspective, in that the inheritance from basic prototypes could be considered a form of cognitive appraisal. Shaver's prototype emotional categories are very similar to Ekman's basic emotional categories (Ekman 1993)(2.1.1). While there is no definitive list of these basic emotions, various researchers have compiled their own lists, with a general agreement on six main primary emotions: fear, anger, happiness, sadness, surprise and disgust. Cowie and Cornelius and Ortony and Turner detail a list of emotional categories compiled from several leading commentators, suggesting an element of cross-pollination and consensus (Ortony 1990; Cowie and Cornelius 2003) (see Appendix A for the full tables). The two tables suggest a consensus regarding the existence of four key emotions: anger, fear, sadness and happiness with a slightly weaker consensus regarding disgust and anxiety. A summary of the emotional categories compiled by Ortony and Cowie is given in Table 1. Although discrepancies do exist e.g. Banse & Scherer differentiate between hot and cold anger (Banse and Scherer 1996) there can said to be some form of general consensus with regard several of the main emotional states.

	Anger	Fear	Sadness	Happiness	Disgust	Anxiety
Lazarus (1999)	✓	✓	✓	✓	✓	✓
Ekman (1999)	✓	✓	✓	✓	✓	X
Buck (1999)	✓	✓	✓	✓	✓	✓
Lewis & Harvard (1993)	✓	✓	✓	✓	✓	✓
Banse & Scherer (1996)	✓	✓	✓	✓	✓	✓
Cowie et al.	✓	✓	✓	✓	X	✓
Arnold (1960)	✓	✓	✓	X	X	X

Ekman, Friesen & Ellsworth (1982)	✓	✓	✓	✓	✓	X
Frijda (1986)	X	X	✓	✓	X	X
Gray (1982)	✓	✓	X	✓	X	✓
Izard (1871)	✓	✓	X	✓	✓	X
James (1884)	✓	✓	✓	X	X	X
McDougall (1926)	✓	✓	X	✓	✓	X
Mower (1960)	X	X	X	✓	X	X
Oatley & Johnson-Laird (1987)	✓	X	✓	✓	X	✓
Panskepp (1982)	✓	✓	X	X	X	✓
Plutchik (1980)	✓	✓	✓	✓	✓	
Tomkins (1984)	✓	✓	X	✓	✓	✓
Watson (1930)	✓	✓	X	X	X	X
Weiner & Graham (1984)	X	X	✓	✓	X	X

Table 1: A summary of the emotions with the most consensus across the literature as detailed by Ortony and Cowie and Cornelius et al. (Ortony 1990; Cowie and Cornelius 2003). The black outline section are the findings from Cowie and Cornelius and the thicker grey outlined section are the findings from Ortony and Turner.

The discrepancies between the lists compiled by the various commentators illustrate the significant problems in determining a set of primary or full-blown emotions, in that descriptive terms may have a different meaning for different commentators. Furthermore, some commentators discuss ‘anger’ and ‘fear’ while others discuss ‘rage’ and ‘afraid’. Can it be taken that ‘rage’ and ‘anger’ or ‘fear’ and ‘afraid’ mean the same thing? While Ekmans research suggests that emotions are universal, some research suggests otherwise (Jack, Blais et al. 2009). Ekmans theories have enjoyed wide support and underpin the idea that there is a basic or primary set of emotions, often taken to be his original six (happiness, sadness, fear, disgust, anger and surprise) with these six, or some of them, often being used in studies examining the acoustic parameters of emotional speech (chapter 3) (Ekman 1987). However, recent research by Matsumoto suggests that, while the initial emotional displays are universal, they quickly modulate into culturally influenced expressions, inline with the views of the Darwinian and social constructivist perspectives (2.1.1 and 2.1.4) (Matsumoto, Willingham et al. 2009). The notion that there is a universal set of basic emotions does not necessarily mean that their exact definition is universally agreed upon. Perhaps this is not possible: as Shiota argues, emotion terms do not relate directly to individual emotional states as: “*..emotion words can encode many features of an*

emotion episode other than internal experience, such as perspectives of the speaker and listener…." (Shiota 2005).

Sabini and Silver make the point that language and its lexicon are culturally relative, with some cultures having words or descriptive phrases for certain emotions while others do not (Sabini and Silver 2005). The example is given of the nonexistence of a word for disgust in the Polish language and the absence of a word for sadness among Tahitians, noting that this is taken by some to indicate the nonexistence of those emotions within these cultures. Sabini and Silvers argument is that there are more emotion terms then there are emotions: numerous synonyms exist for the same emotional state. Consequently, to avoid the subjective nature of emotion terms and their complex nuanced facets, alternative methods of describing emotional states must be considered.

2.5 Emotional Dimensional Representation

This section considers dimensional emotional models. Using a dimensional model allows for a more objective determination of emotional states through the use of axial coordinates. Circular structures of emotional definition have been suggested as a method of establishing contrast between basic (or prototype) emotional categories (Schlosberg 1941; Plutchik 1980; Russell 1980). Circumplex models allow simple visualisations to be made of contrasting emotional categories whilst also displaying underlying emotions that could conceivably be found between both. This method conforms well to the Darwinian, Jamesian, and Cognitive perspectives on emotion (2.1). The descriptive framework is compatible with all three[7] as dimensional models make recourse to the physiological and cognitive aspects of emotional states through the use of activation and evaluation dimensions.

2.5.1 The Development Of Dimensional Models

Wilhelm Wundt saw the potential of an emotional dimensional model, proposing three scales of opposites: pleasure-displeasure, excitation-depression and tension-release (Wundt 1896)[8] . Waston believed that there were only three emotional reactions: love, fear and rage; he recognised that there may be verbal confusion in the use of these terms and therefore proposed labelling them as X, Y and Z (Watson 1929; Strongman, K.T. 2003). While he did not put forward a dimensional model he did recognise the problem with using subjective terms and the advantages in using a more objective approach. Similarly, Millenson built upon Watson's XYZ approach with a three-dimensional model based mainly around emotional intensity and firmly rooted in the primary/secondary emotional tradition (Millenson 1967). Scholsberg used a continuum with six categories (love, surprise, fear, anger, disgust and contempt) arranged along it for users to judge facial expressions; based upon the results he suggested a circular, or possibly an elliptical, bi-polar, two dimensional (pleasantness-unpleasantness and attention-rejection) model (Scholsberg 1941). Scholsberg implemented this model in a series of tests and found that reasonably good

[7] The fourth perspective, Social Constructivist, is also compatible if one accepts its encompassing of the other three

[8] As read in (Schröder 2004b)

predictions of results obtained in his earlier experiments were possible with it (Schlosberg 1952) [
Figure 6 (A)]. He later augmented this with a suggested third axis for activation as he felt that this was better suited to representing emotional states (Schlosberg 1954).

(A) (B)

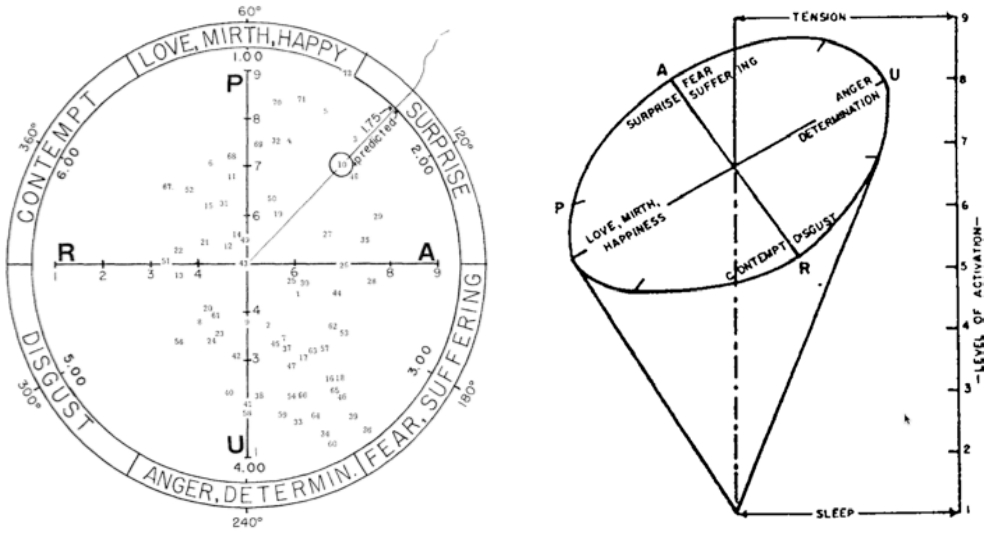


Figure 6: (A) Schlosberg's emotional dimensional model taken from (Schlosberg 1952). Emotional labels are placed along the circumference. (B) Scholsberg's three-dimensional model: an augmented version of his original model with activation as the third dimension, taken from (Schlosberg 1954) .

Work by Osgood, Suci et al and Mehrabian and Russell also suggests that a three dimensional model is necessary to represent emotional states (Osgood, Suci et al. 1957; Mehrabian 1974). Osgood determined that three dimensions were necessary for the description of natural language emotion terms. He termed these dimensions evaluation, activation and potency. However, Osgood later reduced this to two dimensions of pleasantness and activation in relation to the meaning of five facial expressions (Osgood 1960). Mehrabian and Russell originally also contended that three dimensions (pleasure, arousal and dominance) were necessary to describe emotional states.

As with Osgood, Russell later proposed a simpler two-dimensional circumplex model consisting of two axes of pleasure and arousal. This was based upon two types of evidence examined by Russell: laymen conceptualisations regarding emotional states and multivariate analyses of self reported emotional states (Russell 1980). Russell also found that certain emotional terms tended to be situated around the perimeter of his circular model, with emotional states of a lesser intensity lying nearer the middle of the model; he also believed that the circumplex model reflected the layperson's mental mapping of emotional states. Similarly, Abelson and Sermat found that a two dimensional model was adequate for describing emotional facial expressions; like Scholsberg they used pleasant-unpleasantness and tension-sleep (or a slight variation) (Abelson 1962).

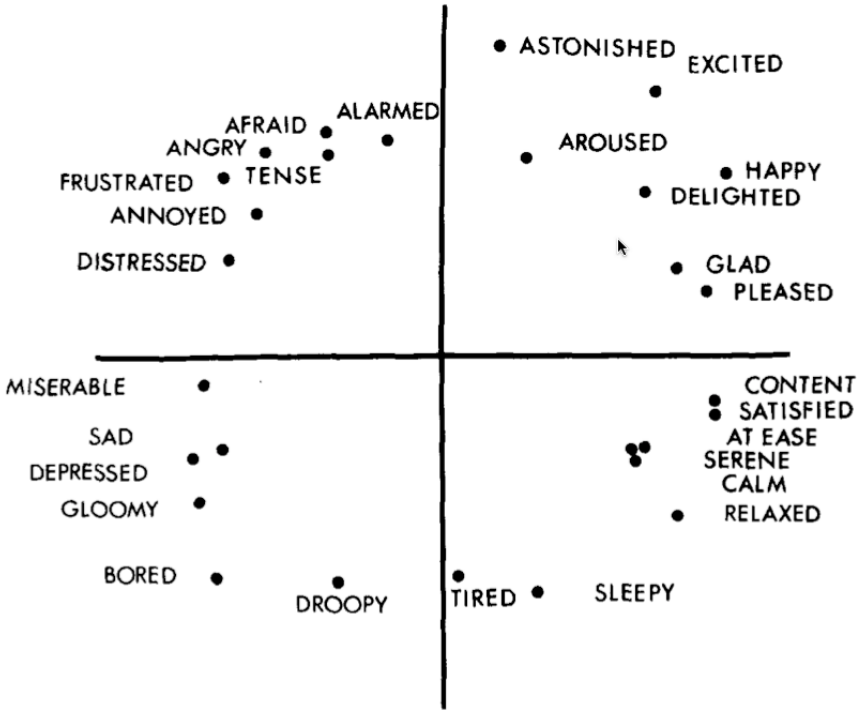


Figure 7: Example of Russell's two-dimensional circumplex model with emotional terms placed within it as determined by Russell's experiments. Taken from (Russell 1980). The vertical axis represents the level of arousal from low (bottom) to high (top), and the horizontal the level of pleasure from unpleasant (left) to pleasant (right).

Plutchik advocated a three dimensional model of intensity, similarity and polarity. Plutchik's model contained eight basic emotions and eight mixed emotions, with all having a biological basis (Plutchik 2001) (2.1.1). Similar emotions are placed beside each other with their polar opposites on the other side of the circle; intensity decreases as one moves down the intensity axis. Colours are used to represent different degrees of the eight basic emotions with secondary emotional states between the primary emotions being mixtures of the primary emotional colours.

(a) (b)

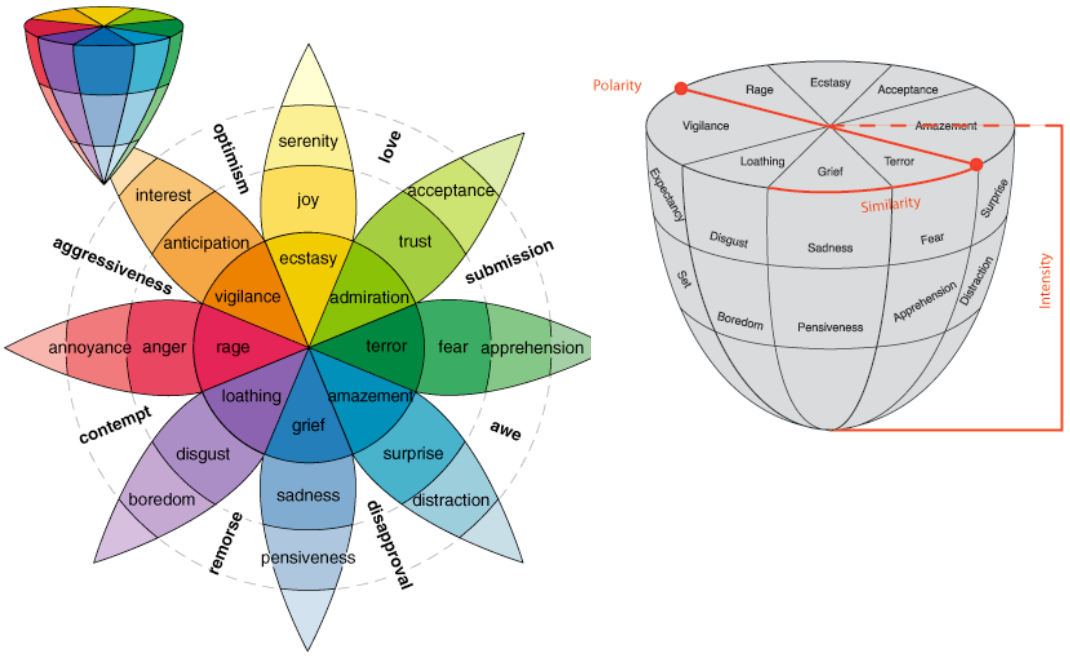
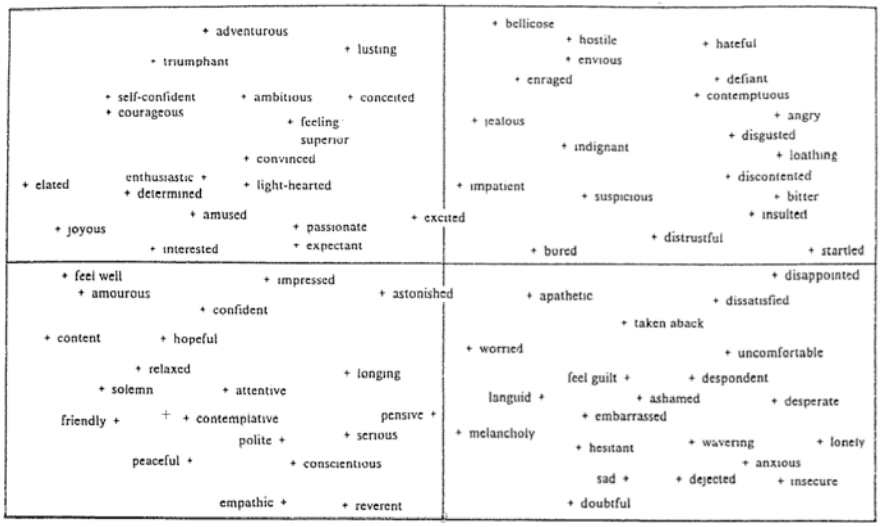


Figure 8: Plutchik's three-dimensional emotional model (a) showing the colour coding and the layout of the emotions. Taken from (Plutchik 2001) and (b) showing the model with the intensity vertical dimension, adapted from (Strongman 2003).

Scherer advocated a two-dimensional model, similar to Russel's, of evaluation and activation, based upon the results of a multidimensional scaling experiment. Scherer found that the distribution of the labels within the model supported the predications of his component process model (2.1.3) (Figure 9a) (Scherer 1984b). Scherer later suggested a three-dimensional tetrahedral affective space, composed of the dimensions of hedonic valence, activation, and control/power, to link appraisal theory to models of affect (Scherer, Dan et al. 2006). Other work by Scherer has suggested that two-dimensions may be inadequate in some situations, arguing that a four-dimensional model of emotion is necessary to represent the six widely shared

theoretical conceptualisations of emotions: appraisals, bodily changes, motor expressions, action tendencies, subjective experiences and emotion regulation (2.5.2) (Fontaine, Scherer et al. 2007).

(a)



(b)

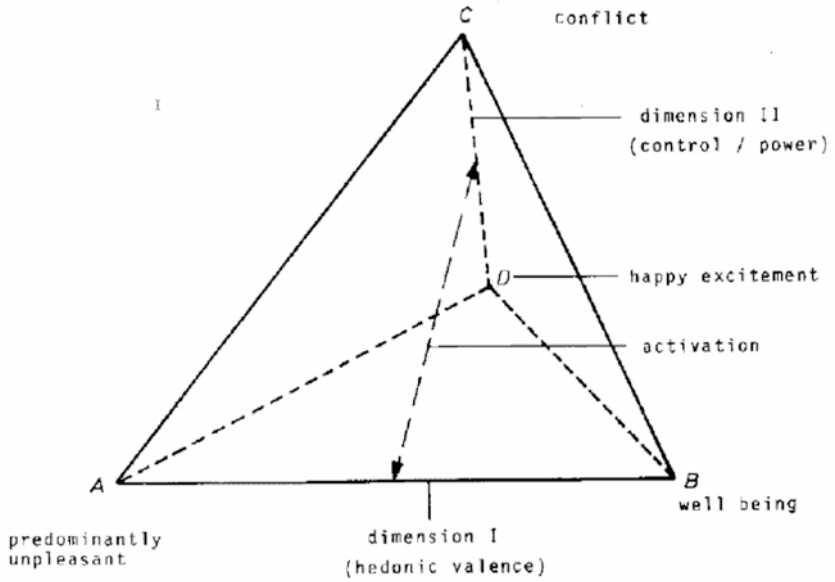


Figure 9: Two different models proposed by Scherer: (a) a two-dimensional model with emotion terms placed within it according to a multi-scaling experiment. The vertical dimension represents level of activity from low (bottom) to high (top), and positive (left) to negative (right) evaluation, taken from (Scherer 1984b) and (b) a tetrahedral model of hedonic valence, activation and control/power taken from (Scherer, Dan et al. 2006).

Cowie et al also describe emotions using two-dimensions of evaluation and activation, culminating in the creation and use of the FeelTrace tool (Cowie, Douglas-Cowie et

al. 2000; Cowie, Douglas-Cowie et al. 2001). The FeelTrace tool allows emotional content to be rated quickly and intuitively by listeners, providing a simple and effective platform for emotional definition. Using such dimensions for the assessment of emotional content is an efficient method: emotional states are graded by how active they are and whether they are positive or negative. The activation dimension is arguably compatible with the biological activation aspects of the Darwinian and Jamesian perspectives (2.1.1 - 2.1.3 and 2.2.2) while the evaluation dimension conforms to the appraisal process central to the Cognitive perspective (Cowie, Douglas-Cowie et al. 2001; Schröder 2004b). The FeelTrace tool uses a colour coding system relative to the position of the cursor within the two-dimensional space that is derived from Plutchik; red is used to signify the most negative evaluation, moving to green for the most positive evaluation; yellow is used to signify the most active state, moving to blue for the least active state (Plutchik 1980): The colours blend depending on the location of the cursor in the dimensional space and are augmented with emotion words placed around the periphery of the model. These words represent the archetypical emotions associated with certain sections of the model and are derived from the prototypes of Plutchik and Russell (Plutchik 1994; Russell 1997; Cowie, Douglas-Cowie et al. 2000).

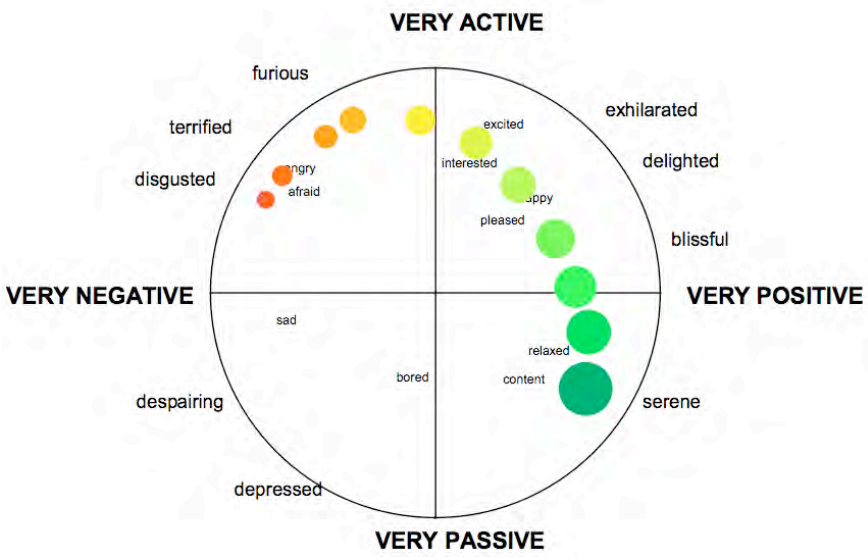


Figure 10: The FeelTrace tool. The cursor changes colour depending on what quadrant it is in. Taken from (Cowie, Douglas-Cowie et al. 2000)

Similar to the FeelTrace tool is the Geneva emotion wheel (Figure 11) (Tran 2004; Banziger, Tran et al. 2005), which also uses colour and emotion words. 16 emotion

families are arranged in the model, each with four different levels of intensity: the larger the circle for each family, the greater the intensity due to its distance from the neutral centre. However, the dimensions are slightly different to the FeelTrace model: the vertical axis represents perceived control[9] from low (at the bottom of the model) to high (at the top of the model) while the horizontal axis represents valence.

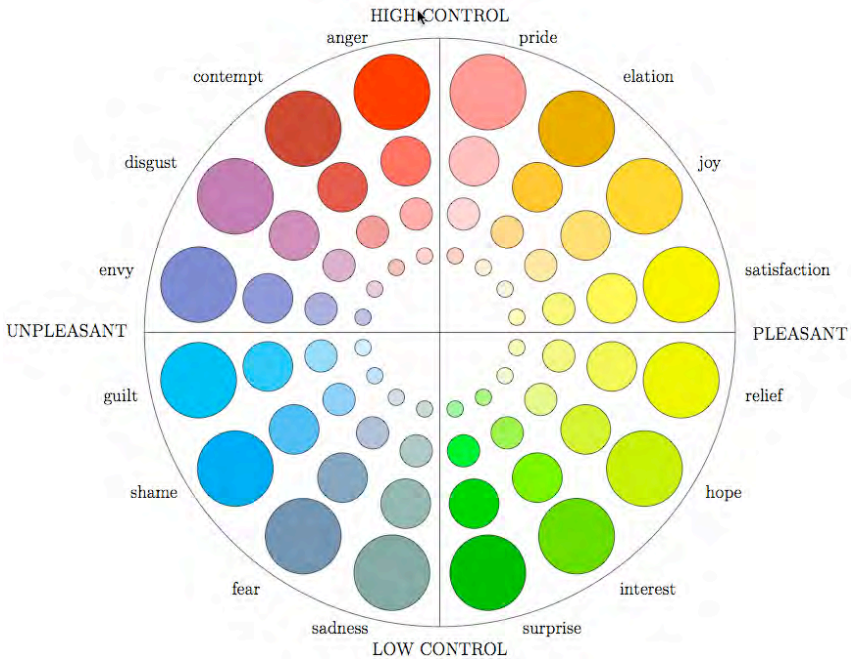


Figure 11: The Geneva emotion wheel. There are 16 emotion categories with four levels of activation for each. The vertical axis represents perceived control and the horizontal axis valence. Activation is represented by size and distance: the further away a level in an emotion category is from the centre, the more active it is. Taken from (Tran 2004).

The organisation of the emotion categories in the model has been experimentally validated and demonstrated that the overall spatial structure of the 16 emotion categories reflected both their similarity ratings and their ratings along the two axes of valence and control (Tran 2004; Banziger, Tran et al. 2005; Steidl 2009).

2.5.2 The Application of Dimensional Models

Dimensional models have been widely used in various studies. Theodoros used a two-dimensional emotional model of arousal and valence to rate speech data from more than 30 films, finding that the dimensional model used was a good representation of the emotions in the speech data; the low level of disagreement

[9] A measure of the perceived control someone has over their emotional state.

among users regarding the ratings of the speech samples was taken as a validation of the rating method (Theodoros 2009). Yang, Lin et al. also used a two-dimensional, arousal-valence model in developing an emotion-based music retrieval system; while they did not comment on the effectiveness of the model, it was used by 99 subjects to annotate the emotional-content of 60 songs (Yang, Lin et al. 2009). Sánchez and Hernández et al. used a two-dimensional model of arousal and pleasure, based on Russell's circumplex model, to convey emotion in an Instant Messaging (IM) service (Sánchez, Hernández et al. 2006). They combined the circumplex model with Ekmans Facial Action Coding System (FACS) to create an IM system they referred to as Russkman IM (Ekman and Rosenberg 2005). This system used emoticons in an affective space to convey emotion quickly and easily within the IM system. Users of the system found it was an easier method of expressing emotion in a simplified manner. Originally the authors found that presenting the adapted Russell circumplex model, complete with emotion terms in each quadrant, overburdened users by requiring too much time and effort to use (Sánchez, Hernández et al. 2006). This is an important consideration in relation to the development of an emotional dimensional rating tool (6.4).

Hanjalic and Xu used a dimensional model to rate the affective content of video clips as part of an automatic feature extraction method (Hanjalic, Xu et al. 2005). They considered a three-dimensional model of arousal, valence and control but concluded that a two-dimensional model consisting of arousal and valence was adequate, citing research by Greenwald which argued that the arousal and valence dimensions can account for most emotional variance and that emotional elicitation using movies, sounds and computers can be mapped onto this dimensional space (Greenwald, Cook et al. 1989; Hanjalic, Xu et al. 2005). They argue, in conjunction with Dietz and Lang, that there was a parabolic shape to the mapping of emotions on the model due to there being very few stimuli that would cause an emotional state consisting of high arousal and neutral valence or high valence and low arousal[10] (Dietz and Lang 1999; Hanjalic, Xu et al. 2005). Laukka and Juslin used a four dimensional model of activation,

[10] Dietz and Lang determined this from scatter plots of responses to the International Affective Picture System (IAPS) and the International Affective Digitized Sounds system (IADS).

valence, potency and intensity, believing that two dimensional models do not allow for the discrimination of certain emotional states such as fear and anger: both have been found to be highly active and unpleasant and so need a third dimension in order to differentiate between them (Laukka, Juslin et al. 2005). To this end they suggested potency as a possible third dimension, considering it as being representative of the cognitive appraisal of an individuals ability to cope with a given situation; it is also sometimes referred to as dominance, power or control (Laukka, Juslin et al. 2005). The fourth dimension, intensity, was used to capture the strength of the emotional state: the Geneva Emotion Wheel uses the distance from the neutral centre as a measure of intensity, as does the FeelTrace tool in conjunction with a blending of colours (2.5). Russell's early circumplex model also represents intensity using distance from the centre of the model: more intense emotional terms lie nearer the periphery while those of a lesser intensity lie nearer the centre (Russell 1980) (2.5).

Like Laukka and Juslin, Fontaine and Scherer also argue for a four-dimensional model, consisting of evaluation-pleasantness, potency-control, activation-arousal and unpredictability (Fontaine, Scherer et al. 2007). The four-dimensional model was derived from a web-based study (across three languages: English, Dutch and French): each participant was given four emotional categories/terms chosen randomly from a set of 24. They were then asked to rate each term in relation to 144[11] emotion features using a nine point Likert scale. These emotion features are a deployment of the six emotion components that Fontaine and Scherer believe to be widely shared theoretical conceptualisations of emotions: appraisals, bodily changes, motor expressions, action tendencies, subjective experiences and emotion regulation. Conversely, Fontaine and Scherer do contend that their study is more about the perceived meanings of emotional terms and a wider decontextualised emotional conceptualisation, as opposed to the more episodic and experiential aspects of emotion. Thus they maintain that their findings may not represent the dimensions of emotional experience: they do believe, however, that the dimensions of arousal and valence are important aspects of emotional experience and that the most expedient dimensional model to be used depends on the research and the questions being asked (Fontaine, Scherer et al. 2007).

[11] For the complete list of 144 items used, see (Fontaine, Scherer et al. 2007)

2.5.3 Dimensional Ratings: Considerations

Considering the findings of the previous section, it is apparent that two or three dimensions are often used in emotional dimensional models; these models aim to capture the salient physiological and neurological aspects of emotional experience via the dimensional representation of biological activation and cognitive appraisal. Arguably this a more objective approach than the use of subjective emotion terms as discussed in the previous sections (2.3 and 2.4) and goes some way in avoiding the numerous problems associated with their use. However, it has been argued that dimensional models do not capture the full range and complexity of emotional states. Lazarus in particular found them to be too reductionist and believed that they did not account for the communicative and expressive role of emotion, which is the focus of Lazarus' appraisal centric theories of emotion (Lazarus 1991). Likewise, Fontaine and Scherer have argued for a four-dimensional model (2.5.2) (Fontaine, Scherer et al. 2007). Conversely, Russell believes that a circumplex model adequately encapsulates the layman's mental mapping of emotion, while Cowie et al. see dimensional models as being linked to the emotional experience itself (Russell 1980; Cowie, Douglas-Cowie et al. 2001).

Though dimensional models may be somewhat reductionist, their widespread use suggests that they can be used to represent important aspects of emotional experience and their use does not preclude their integration into a wider, more descriptive emotional framework (Schröder 2004b). Two-dimensional models have been successfully utilised in a variety of studies (Hanjalic, Xu et al. 2005; Sánchez, Hernández et al. 2006; Oehme 2007; Theodoros 2009; Yang, Lin et al. 2009) (2.5.2). Research has also suggested a strong correlation between the activation dimension and certain acoustic parameters (Pereira 2000; Schröder 2004a; Schröder 2004b; Laukka, Juslin et al. 2005) (3.6 and chapter 9). Dimensional models arguably offer a better alternative than adopting a set of inconclusive emotional terms and are well suited for use in listening tests (6.4.2). Russell maintains that using a limited set of emotional labels/categories means that listeners/labellers are forced into using labels that they might not agree with; work by Russell has shown that different forced-choice formats resulted in different labels being applied to the same stimuli (Russell 1993; Russell 1994; Russell, Bachorowski et al. 2003). Using a dimensional model as

part of a listening test allow users to make dimension based perceptual judgements as opposed to a forced-choice discrimination between categories. Furthermore emotional dimensions are capable of representing high intensity primary emotional states and low intensity underlying emotional states (Schröder 2004a).

2.6 Emotion, Mood and Affect

Mention must be made of the differentiation of the terms emotion, affect, attitude and mood. Often the literature will use the terms emotion and affect interchangeably as synonyms of each other. However, as Sloman points out, the two have different meanings: affect covers a wider range of concepts such as moods, attitudes, desires, intentions and dislikes, while emotion is a special case of the more general concept of affect (Sloman 2005). Similarly, Wichmann notes that attitude and emotion are often conflated, being used differently by psychologists and linguists (Wichmann 2000). For psychologists, attitude is a term used to describe the intention to act and is often used as an all-encompassing term for a complex system that includes emotions. Gobl and Ni Chasaide et al. accept that there is a distinction to be made between emotion, mood and attitude with all three being part of a wider set of affective attributes (Gobl and Ní Chasaide 2000). Beedie at al. carried out a comprehensive investigation and review of the literature regarding the distinction between mood and emotion (Beedie 2005). Mood was conceived as being a ‘background’ feeling state with no specific direction or cause while emotion had a cause and a direction. One of the main distinguishing features between mood and emotion was duration: mood was considered a long-term state while emotion was regarded as shorter and more spontaneous. Parkinson sees emotions as being temporally specific events, whereas moods develop over a period of time, usually due to a culmination of minor events (Parkinson 1996). Frijda distinguishes between emotion and mood by arguing that emotion is directly caused by something, we feel sad about an event, we feel angry with someone; conversely, mood is indirectly caused by something and is more of a general disposition e.g. being in a good mood and having a general positive disposition rather than it being directed at any one thing or event (Frijda 1994; Brave and Nass 2008).

Malatesta and Murray et al. see the term emotion as referring to brief and intense experiences, and mood as being of a lower intensity but with a more prolonged duration (Malatesta, Murray et al. 2009). As with some of the definitions already discussed, Scherer sees emotion as being episodic in nature, defining it as: *"…an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism"*(Scherer 2001a), while mood is characterised by low intensity with a long duration and no direct or apparent cause (Scherer 2000e). A consensus is evident regarding emotions as being brief and more intense than moods. Considering the discussion in 2.3 and 2.4, primary and full-blown emotions are characterised by their intensity, with underlying emotional states being less intense than primary/full-blown emotions but possibly more intense than moods. While the distinction between underlying emotional states and moods with regard to intensity may not always be obvious, the directed nature of underlying emotions is still a distinguishing factor.

Cowie and Corenelius use the term 'emotional states' to refer to all states involving emotion, primary or otherwise, as opposed to mood or affect, and this use of the term is adopted in this document[12] (Cowie and Cornelius 2003).

2.7 Conclusions

This chapter has considered various approaches to the definition and examination of emotion. The four prominent psychological perspectives were first examined, observing that some researchers firmly root emotions in biological processes, while others view them as the outcome of cognitive appraisals or as socially constructed appraisal processes (2.1). The four perspectives are not mutually exclusive and can be considered different aspects of a wider emotional theory; Cornelius sees the first three perspectives as being on a path of convergence and the social constructivist perspective as possibly encompassing all three (2.1.5). Following this, examination of

[12] The term affect is used later in this document in examining the literature on the acoustic correlates of emotional speech (chapter **Error! Reference source not found.**), as it used extensively in the literature being reviewed.

the literature suggested that there are important physiological and neurological facets to emotion (2.2). Subsequently, the idea of full-blown and primary emotions was discussed, arguing that they are often seen as prototypical emotions with underlying and secondary emotions deriving from them (2.3 and 2.4). However, there is disagreement regarding the full list of both primary and secondary emotions: lists compiled by several commentators differ, agreeing only on a small sub-set of emotional terms and it was argued that underlying emotional states are of greater interest due to their constituent part in the human communicative process (2.3, 2.4) (Table 1). The dimensional representation of emotions was then considered as a method of avoiding the use of subjective emotion terms (2.5). Dimensional models have been widely used by numerous researchers and possibly capture important aspects of emotional experience. Finally, the difference between affect, affective states, attitude, mood and emotion was examined, determining that there was a distinction to be made between these terms despite them often being used synonymously in the literature (2.6). The review of the literature in this chapter has given rise to the following research question**:**

RQ1: Is a two-dimensional model adequate to capture some salient aspects of natural underlying emotional speech?

3. Acoustic Parameters of Emotional Speech

This chapter discusses various acoustic parameters of speech that may indicate the presence of certain emotional states. One of the aims of this research is to examine the acoustic correlates of natural underlying emotional speech. Therefore aspects of the speech signal that may indicate emotional dimensions are considered. The main measurable elements of the speech signal that will be examined are collectively known as prosody and include fundamental frequency (F0), intensity, duration, amplitude (intensity) dynamics and voice quality. However, the term prosody can refer to various different speech properties, depending on the approach taken. The three main representations of prosody are discussed, examining the differences in the meaning of the term prosody as it relates to different disciplines (3.1). Voice-quality is considered separately in order to better define the term prosody and limit its use to a few specific acoustically measurable elements (3.4).

A review of the literature of the acoustic correlates of emotional speech is carried out to determine the acoustic parameters found to be related to certain emotional states (3.3). The majority of the studies examine a relatively small group of primary emotional states and their acoustic realisations in speech, though a small number of studies have examined acoustic correlates in relation to emotional dimensions (3.6).

3.1 Prosody

The term prosody is a complex term referred to by numerous researchers from different disciplines within the Speech and Language paradigm (Cutler, Dahan et al. 1997) and as a result it is difficult to make a distinction between the form and the function of prosody (Mixdorff 2002). Some researchers tend to view prosody as being an abstract component of speech with little regard for its actual acoustic realisation. Other researchers focus on the acoustic features and tend to see prosody as a synonym of suprasegmental features of speech (Werner and Keller 1994), which include elements such as pitch, loudness (intensity) and speech rate. Cutler believes that the majority of definitions of prosody are a mixture of the abstract and the acoustic realisation (Cutler, Dahan et al. 1997). The idea that prosody is a synonym for suprasegmental features is supported by Werner and Keller (Werner and Keller 1994),

based on a classical study on prosody by Lehiste (Lehiste 1970). Dutoit also views prosody as referring to suprasegmental features such as pitch, loudness and syllable length (Dutoit 1997). Dutoit distinguishes between 3 different representations of prosody as outlined in Table 2:

Acoustic	Perceptual	Linguistic
Fundamental Frequency (F0)	Pitch	Tone, intonation, Aspect of stress
Amplitude, Energy, Intensity	Loudness	Aspect of stress
Duration	Length	Aspect of stress
Amplitude dynamics	Strength	Aspect of stress

Table 2: Dutoit's three representations of prosody with their respective properties. The acoustic reprsentation describes measurable acoustic parameters related to sound and speech. The perceptual representation is synonomous with the acoustic representation, referring to the same parameters using different terms. The linguistic representation is the least descriptive, referring to numerous acoustic parameters as aspects of stress (Dutoit 1997).

The acoustic representation refers to measurable acoustic properties of the speech signal. Within this representation, segment duration is also included despite not being an entirely acoustic feature. Duration can be applied to various aspects of the acoustic level: vowel duration, pitch duration, intensity duration etc. Even within the acoustic representation of prosody, studies can differ with some focusing on fundamental frequency (F0) while others focus on temporal features (duration, speed of delivery) along with F0. Furthermore, acoustic prosodic variables are not limited to one particular level of speech but can operate from something as small as a vowel to an entire utterance, thus making prosodic measurements context sensitive; for example, the intensity of an entire utterance can be considered or further broken down into smaller discreet intensity measurements at particular points in time.

As Table 2 shows, the properties of prosody at the acoustic level correspond to prosodic properties in the perceptual and linguistic representations. The perceptual representation refers to how the acoustic factors of prosody are perceived by the

human perceptual system, with acoustic factors such as intensity being perceived as loudness. The linguistic representation refers to the F0/pitch as intonation while the other prosodic factors are taken to be aspects of stress. The three representations as described by Dutoit are useful in decoding the meaning of the terminology associated with prosody as there can be some confusion about the various terms used within the different representations (Kochanski 2006). Understanding the terminology used within the different approaches allows research to focus on one particular prosodic representation while being aware of how this focus relates to the other representations. This thesis is primarily concerned with the acoustic representation of prosody and as such focuses on the acoustic representation/definition of prosody.

A review of the literature relating to the acoustic parameters of emotional speech is next considered. While not all studies examine the same acoustic parameters (some use more than others), they are the most commonly measured prosodic parameters. While not an exhaustive review, it is comprehensive enough to demonstrate that certain acoustic parameters are worth investigating in relation to underlying emotion in speech. As this chapter will show, despite different methodologies and goals, a consensus of sorts can be seen to emerge. Rather than focus solely on a small group of singular studies, studies and reviews that themselves examined a range of studies are considered in order to obtain a broader picture of the findings concerning the acoustic parameters of emotional speech.

3.2 Defining The Acoustic Parameters

The section discusses the methodology of various studies before discussing their findings under a number of different headings: pitch, intensity, pitch contour, speech rate and voice quality. These parameters are the most widely examined in relation to emotional speech, though other parameters are often considered alongside them (Spackman, Brown et al. 2009). Some of these terms are briefly defined and examined in the following sections.

3.2.1 Pitch

Sound is defined as the displacement of air molecules that correspond to the original disturbance, with the displacement of the air molecules occurring in the direction that

the disturbance is travelling (Pohlmann 2000). Three basic quantities are used to describe a waveform: frequency, amplitude and phase. The frequency and amplitude aspects are of most interest here.

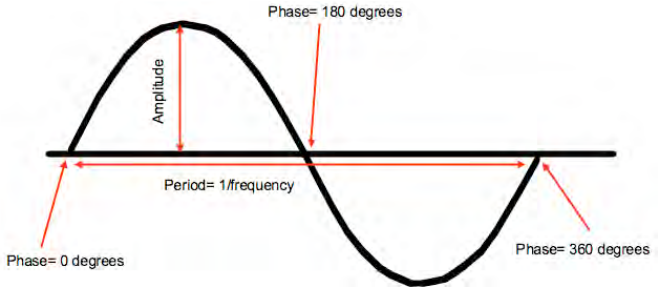


Figure 12: Diagram of a sine wave illustrating the three component aspects of sound: amplitude, frequency and phase. Adapted from (Cook 2001).

As discussed in 3.1, pitch is the perceptual realisation of the frequency acoustic parameter. Frequency is defined as the number of times that a waveform repeats itself in 1 second and is measured in Hertz (Hz); the calculation of pitch in Hz is carried out in the frequency domain, with the mean pitch of a speech segment being calculated over the duration of the segment. A 200Hz waveform repeats itself 200 times a second, while a 20Hz sound will repeat itself 20 times a second; the frequency of a sound is a measure of the periodicity of its waveform and the frequency range of human hearing is 20Hz to 20KHz (Rumsey and McCormick 2002). Speech is made up of periodic and aperiodic sounds, with the periodic parts of speech signal having a perceptual and measurable frequency (Howard and Angus 1999). The periodic portion of the speech signal can thus be measured and plotted to give a pitch contour which is a continuous line that represents the variation in pitch over a given amount of time (Dutoit 1997) (Figure 13).

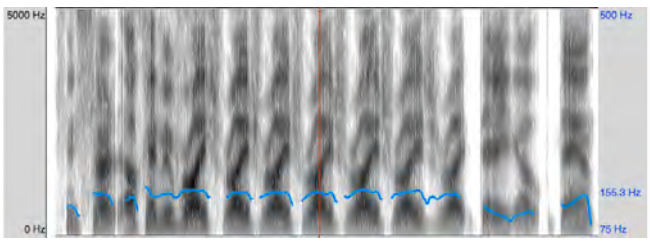


Figure 13: Example of a pitch contour for a speech segment. The blue line is the contour overlaid on the spectrogram of the speech segment.

Breaks in a pitch contour indicate aperiodic portions of the speech signal. Ververidis and Kotropoulos found the literature they reviewed to contain very little in the way of a systematic description of pitch contour shape (Ververidis and Kotropoulos 2006).

While the human hearing system can perceive a huge amount of frequencies (20Hz-20KHz), it has a non-linear frequency response. As a result, a psychoacoustic logarithmic scale is often a better model of the human hearing systems pitch perception than the linear Hertz scale. There are numerous psychoacoustic scales: mels, Bark, ERB-rate and semitone scale; the semitone and ERB-rate best reflect the frequency response of the human hearing system (Nolan 2003). This issue is discussed further in 9.1.1, where pitch values are reported in Hertz and semitones.

3.2.2 Intensity/Amplitude

Amplitude is a measure of the amount of pressure exerted by a waveform and the resulting displacement of a medium as it moves through it. More precisely, amplitude is the amount of energy delivered by a sound source to a particular point, defined as $watts/m^2$ (Roederer 2008). The more energy used to generate a sound, the greater the amplitude and perceived loudness. Since the human auditory system is able to detect a huge range of intensities, the more convenient Sound Intensity Level (SIL) and Sound Pressure Level (SPL) values are used instead of $watts/m^2$, the unit of both being the bel and more the commonly used decibel (dB). SIL and SPL are both logarithmic scales: SIL is a logarithmic ratio value with reference to the threshold of hearing (defined as $10^{-12} Watts/m^2$) and is defined as:

$$SIL = 10\log_{10}\left(\frac{I_{actual}}{I_{ref}}\right)$$

Equation 1: Formula for calculating Sound Intensity Level.

Where I_{actual} = the actual sound power density (Wm^{-2}) and I_{ref} = the threshold of hearing (Howard and Angus 1999). While sound intensity level is a useful measurement, SPL is the more common quantity when describing amplitude, and also makes reference to the threshold of hearing ($20\,\mu Pa$), and is defined as:

$$SPL = 20\log_{10}\left(\frac{P_{actual}}{P_{ref}}\right)$$

Equation 2: Formula for calculating Sound Pressure Level.

Where P_{actual} = the actual pressure level (Pa) and P_{ref} = the threshold of hearing (Howard and Angus 1999).

3.2.3 Speech rate

Speech rate is one method of measuring how fast a person is speaking and is usually measured in words or syllables per second (Laver 1994; De Jong and Wempe 2009). It is calculated by:

$$SR = \frac{s}{t}$$

Equation 3: Equation for calculating Speech Rate.

Where s = number of syllables in an utterance and t = duration of the utterance.

In speech rate, the duration of an utterance is the total length of the utterance, including pauses. In contrast, articulation rate is calculated over the length of an utterance excluding pauses and is a measure of how fast syllables are produced (Trouvain and Barry 2000). An increase in speech rate can be indicative of a shortening of the pauses in the speech utterance, while an increase in articulation rate indicates an increase in the speed that syllables are produced and is often considered a better measure of how fast a person is actually speaking. The findings regarding speech rate in literature reviewed are more consistent than those for articulation rate.

3.3 The Acoustic Correlates of Five Primary Emotional States

Five emotional states are examined in relation to each of the five acoustic parameters: anger, fear, happiness, sadness, disgust. While other emotional states are examined in conjunction with these five, they are the most commonly examined in the literature, as they are considered primary/full-blown emotional states (2.3 and 2.4).

Ververidis and Kotropoulos (Ververidis 2006) carried out an overview of emotional speech recognition, looking at existing emotional data sets as well as the most common acoustic parameters used for emotional speech recognition. They examined 14 studies relating to emotional cues in speech with regard to five emotions: anger, fear, joy, sadness and disgust. These studies considered the most frequently used acoustic parameters of pitch mean (in Hz), range, variance and contour, intensity mean and range, speech rate (calculated on pitch pulses or syllables per second) and transmission duration between utterances. The emotional categories examined in the literature they reviewed are similar to the basic emotions that most commentators appear to agree on (2.3 and 2.4). Juslin and Laukka carried out a large meta-analysis of 104 studies into the acoustic parameters of emotion (Juslin and Laukka 2003). They examined the emotion categories used in the 104 studies, the method used to obtain emotional speech and the findings regarding the acoustic parameters of each emotion category. While a number of different emotions were examined across the 104 studies, five primary emotional categories appeared in the majority of them: anger, fear, happiness, sadness, disgust. Juslin and Laukka focused on a slightly different set five of emotions in the latter part of their paper: anger, fear, happiness, sadness and tenderness.

Cahn (Cahn 1990) examined five studies into the acoustic correlates of emotional speech carried out between 1939 and 1974. Using the findings of these studies, she derived a number of prosody rules for the expression of emotion in synthesised speech. These rules were used to modify certain acoustic parameters of a synthesised voice to convey six emotions: glad (happiness), sadness, anger, fear, surprise and disgust. The acoustic parameters modified according to the prosody rules were: fundamental frequency (F0), F0 mean, F0 range, F0 pitch contour, tempo and voice quality (VQ). The veracity of the prosody rules was verified using a perception test, in which five sentences were synthesised for the six emotions and listeners had to identify the emotion in each sentence. In a wide ranging study in to the literature on vocal emotions, Murray and Arnott (Murray 1993) adopted the primary/secondary emotional distinction in assessing the acoustic parameters of emotional speech (2.4). In this case, the primary emotional category consisted of five emotions: anger,

happiness, sadness, fear and disgust, identical to the five emotions examined by the majority of researchers discussed in this chapter. The secondary emotional states they examined were: grief/sorrow, affection/tenderness and sarcasm/irony. Other secondary emotions were mentioned but did not receive as much attention due to their limited mention in the literature reviewed. Murray and Arnott used their findings to develop an expressive synthesised voice using the HAMLET system, concluding that emotional sounding synthetic voices can be produced with a degree of realism (Arnott 1995).

Over the course of several studies, Scherer et al. (Scherer 1972; Scherer 1981a; Scherer 1986b; Scherer, Johnstone et al. 2003) have investigated the acoustic parameters correlated to certain emotional states. Scherer's' findings across the research are relatively consistent, and consider an expanded range of emotional categories that include the main five: anger, fear, happiness, sadness and disgust. Early work by Scherer examined the literature relating to the vocal indicators of emotional states for eleven emotions, including the 'common' five (Scherer 1979; Scherer 1981b), replicating and confirming the findings in subsequent research. Banse and Scherer used professional actors to portray fourteen emotions (totalling 224 emotional portrayals) (Banse and Scherer 1996). These portrayals were judged by listeners and results indicated a better than chance accuracy, reproducing the findings of earlier work, and supporting the predictive powers of the Component Process Model (Scherer 1984b) (2.1.3). Included in the fourteen emotions portrayed in the study were the five common emotions discussed in this chapter. Scherer found that these five emotions were characterised by certain acoustic parameters. The five emotions were compared to a neutral state and it was noted that anger was the most examined emotion in the literature, with some contradictory findings, mainly due to the fact that anger can manifest in a number of different ways. Scherer himself makes the distinction between hot anger and cold anger, with both seemingly having different acoustic parameters (Scherer 1986b; Scherer, Johnstone et al. 2003).

Gobl and Ni Chasaide examined voice quality in relation to certain emotional states (Gobl and Ní Chasaide 2000; Gobl, Bennett et al. 2002; Gobl and Ní Chasaide 2003). In two particular studies, a synthesised voice uttering a phrase in Swedish ('Ja adjö)

was modified to produce seven different voice stimuli: tense, breathy, whispery, creaky, modal (normal synthesis of the utterance), harsh and the experimental lax-creaky voice (Gobl and Ní Chasaide 2000; Gobl, Bennett et al. 2002). Eight listeners listened to ten randomisations of the seven voices, rating them using eight seven-point scales, with opposite affective attributes at each end of the scales: relaxed/stressed, content/angry, friendly/hostile, sad/happy, bored/interested, intimate/formal, timid/confident, and afraid/unafraid. Within these sixteen attributes were four affective attributes similar to the emotion categories analysed by the other researchers: angry, sad, afraid (fear) and happy. O' Reilly and Ni Chasaide also investigated pitch contour in relation to portrayed emotions (O'Reilly and Ní Chasaide 2007). In this study pitch contours from six portrayed emotions were examined (surprised, bored, neutral, angry, happy and sad).

3.3.1 Pitch/F0

Ververidis and Kotropoulos found that speech classed as angry had the highest mean pitch of all the emotional categories examined (Ververidis 2006). Disgusted speech had a low mean pitch, while fearful speech was characterised by a high mean pitch level; most of the research they reviewed reported a wide pitch range for fearful. Sad speech had a low mean pitch, while few studies reviewed gave as much consideration to happy/joyful speech with regards to pitch measurements. In their extensive review, Juslin and Laukka found angry speech was associated with high pitch level and high pitch variability (Juslin and Laukka 2003). Fearful speech also had a high pitch level with only a small amount of pitch variability. Happy speech also had a high pitch level, and a high level of variability. In contrast, sad speech had a low pitch level, and low variability. In her study Cahn found angry speech to have large pitch transitions that generally had a downward inflection, seemingly at odds with the usual reported high-pitch characteristic of angry speech (Cahn 1990). Disgusted speech had a low pitch with a wide pitch range, however in some cases the pitch was reported to be close to, or the same as, the neutral utterance used in the study. Fearful speech was found to be high-pitched with a wide pitch range, while glad speech had a wide pitch range and large pitch variations, with an overall tendency to be high-pitched. Sad speech had very little variation in pitch, a narrow pitch range and, in general, was found to be soft sounding and low-pitched. Surprised speech had little in the way of acoustic descriptors, but was generally found to be high-pitched. Murray and Arnott

found that, while some of the findings in the literature they reviewed regarding angry speech were contradictory, it was generally found to be high-pitched, with an increased average pitch and pitch range (Murray 1993). Likewise, happy speech also had an increased average pitch and range. Sad speech had a lower average pitch and range while fear had an increased average pitch and pitch range. Finally, Murray and Arnott found disgusted speech had a low average pitch, with a slightly wider pitch range, with wide downward directed pitch inflections. Scherer et al. found that angry speech characteristically had an increased pitch with downward directed pitch contours and an increased pitch range (Scherer 1979; Scherer 1981b; Banse and Scherer 1996; Scherer, Johnstone et al. 2003). Fearful speech had an increased mean pitch and pitch range with Scherer noting that there was a good deal of agreement across the literature regarding the acoustic cues of fear. Sad speech had a decreased mean pitch and pitch range with pitch contours that are generally downward in nature. Joyful/happy speech was usually found to have a high pitch mean and pitch range. Some studies found disgusted speech had an increased mean pitch while others found it had a decreased mean pitch.

	Anger	**Fear**	**Sadness**	**Happiness**	**Disgust**
Pitch Mean	Increased /high V&K, Ca, M & A, Sch, J&L	Increased/high, V&K, Ca, M&A, Sch, J&L	Decreased/low, V&K, Ca, M&A, Sch, J&L	Increased/high, V&K, Ca, M&A, Sch, J&L	Decreased, V&K, Ca (or equal to), M&A,
Pitch Range	Increased V&K, Ca, M&A, M, Sch	Increased, V&K, Ca, M&A,Sch	Decreased, V&K, Ca, M&A, Sch	Increased, V&K, Ca, M&A, Sch	Increased for men, decreased for females, V&K. Decreased, Ca. Increased, M&A

LEGEND	Cahn= **Ca**	Scherer et al= **Sch**	Murray & Arnott= M&A	Ververidis & Kotropoulos = **V & K**	Murray Et al= **M**	Gobl & Ni Chasaide = **G&Ni C**

Table 3: A summary of the findings in the literature reviewed regarding pitch mean and range. There is a strong consensus for pitch mean and range for anger, fear, sadness and happiness (highlighted in green), there is a weaker consensus for disgust (highlighted in yellow).

3.3.2 Pitch Contour

As discussed in 3.3.2, pitch contour is a temporal representation of pitch changes, illustrating pitch changes over the duration of an utterance. The literature reports the

general trend of the pitch contours of the speech examined in the various reviewed studies. Ververidis and Kouropoulos findings suggested that the pitch contour slope for fearful speech tended to incline upward, as did the contour slope for sad speech while declining downward for joy (Ververidis and Kotropoulos 2006). They found that pitch contour was important in separating fearful speech from joyous speech; the two have similar acoustic correlates but the pitch contours are different. Juslin and Laukka found angry, fearful and happy speech all had rising pitch contours while sad speech had a falling contour (Juslin and Laukka 2003). As mentioned in the section on pitch (3.3.1) Cahn found angry speech, in general, to have a variable pitch contour with a downward directed slope (Cahn 1990). Cahn's findings regarding glad speech suggest a variable pitch contour with an upward slope. Sad speech had a relatively stable contour with little variation, while surprised speech was generally found to have a rising pitch contour slope. Murray and Arnott found angry speech to have a highly variable, rising contour with abrupt pitch changes on stressed syllables (Murray 1993). They found happy speech to have smooth pitch changes with a general upward inflection. Sad speech had downward inflected pitch changes while fearful speech was described as having normal pitch changes, with no large pitch variations, suggesting a relatively stable contour. Mozziconacci et al. found that there was no one to one relationship between pitch contour and emotion (Mozziconacci 1997). While there is a general lack of agreement regarding the relationship between pitch contour and emotional state, Ni Chasaide and O Reilly found some difference in the shape of the speech pitch contour for five different emotional categories: sad, happy, bored, surprised, angry and a neutral utterance (O'Reilly and Ní Chasaide 2007). While the contours for sad, bored and neutral speech were very similar there was a more noticeable difference between the contours for happy, surprised and angry speech. They also found that certain measured parameters were useful in differentiating between emotions of high and low activation but were not indicative of a singular emotional state.

	Anger	Fear	Sadness	Happiness	Disgust
Pitch Contour	Rising,Ca, J&L. Raised, M. Downward directed, Sch.	Rising, Ca, J&L. Raised, M, Sch	Rising, V&K. Lowered, M. Downward, Sch Falling, J&L	Falling, V&K. Rising, Ca. Raised, M. Rising J&L.	NA

LEGEND	Cahn= **Ca**	Scherer et al= **Sch**	Murray & Arnott= M&A	Ververidis & Kotropoulos = **V & K**	Murray Et al= **M**	Gobl & Ni Chasaide = **G&Ni C**

Table 4: A summary of the findings in the literature reviewed regarding pitch contour with a strong consensus for fear, sadness and happiness. There is a weaker consensus for anger and none for disgust.

3.3.3 Intensity

The findings in the literature regarding emotion and intensity closely mirror the findings in relation to pitch. In the majority of cases there appears to be a positive relationship between the two parameters: emotional speech with a characteristically high mean pitch usually also has a high mean intensity. Ververidis and Kotropoulos found angry speech had an increased intensity, as did fearful speech, while disgusted and sad speech had a low level of intensity (Ververidis 2006). Similar to their findings regarding pitch Juslin and Laukka found angry, and happy speech to have a high intensity level (Juslin and Laukka 2003). Fearful speech was reported as having a low level of intensity except in the case of speech related to panic fear, which was reported as having a high intensity level, while sad speech also had a low level of intensity. Both anger and fear had a high level of intensity variability while sadness had a low level of variability. Cahn described anger as having an increased amount of intensity with quick rises in intensity level (Cahn 1990). Fear and surprise had an increased intensity level, whereas sadness has a slightly decreased intensity level. The intensity level of disgust was unchanged from the neutral level of the study. Glad speech had an increased intensity level, being described as loud and 'blaring'. For Murray and Arnott, anger and happiness had an increased intensity level in contrast to sadness and disgust, who both had a decreased intensity level (Murray 1993). The intensity level of fear was unchanged from the neutral state. Scherer's findings regarding intensity closely mirror his findings in relation to pitch (Scherer 1979; Scherer 1981b; Banse and Scherer 1996; Scherer, Johnstone et al. 2003). Anger, fear,

happiness and disgust are all described as having an increased mean intensity, while sadness has a decreased level.

	Anger	**Fear**	**Sadness**	**Happiness**	**Disgust**
Intensity Mean	Increased, V&K, Ca, M&A, Sch, J&L	Increased or no change, V&K. Low, J&L (except in panic fear). Increased, Ca, Sch. Normal, M&A	Decreased, V&K, Ca, M&A, Sch, J&L	Increased, V&K, Ca, M&A, Sch, J&L (medium to high)	Decreased, V&K. No change, Ca. Increased, Sch.
Intensity Range	Increased, V&K, Sch.	Increased, Sch.	Decreased, V&K, Sch.	Increased, V&K. Increased (Joy) Decrease/no change (Happy), Sch	

LEGEND	Cahn= **Ca**	Scherer et al= **Sch**	Murray & Arnott= M&A	Ververidis & Kotropoulos = **V & K**	Murray Et al= **M**	Gobl & Ni Chasaide = **G&Ni C**

Table 5: A summary of the findings in the literature reviewed regarding intensity mean and range. There is a strong consensus regarding intensity mean for anger, sadness and happiness with a weaker consensus for fear. There is a weak consensus regarding intensity range for anger, sadness and happiness.

3.3.4 Speech Rate/Tempo

As discussed in 3.2.3, speech rate is one method of determining speech rate, the other being articulation rate, but the findings regarding speech rate are more consistent across the reviewed literature. Ververidis and Kotropoulos found disgust and sadness to have slow speech rates, while fear had shorter pauses between speech segments, giving an increased speech rate (Ververidis 2006). Juslin and Laukka established that anger, fear and happiness had fast speech/tempo rates, while sadness had a slow speech/tempo rate (Juslin and Laukka 2003). Cahn found anger to have a fast tempo along with fear, happiness and surprise (Cahn 1990). Sad speech had a slow tempo while the speech rate of disgust was unchanged from that of the neutral utterance. Murray and Arnotts findings regarding speech rate coincide with those of Cahn and Ververidis and Kotropoulos (Murray 1993). The speech rate of anger, happiness and fear was increased in conjunction with their pitch and intensity levels (fear being the exception as it was found to have had no increase in intensity). Disgust and sadness

both had slower speech rates, coinciding with a decrease in pitch and intensity levels (3.3.1 and 3.3.3).

	Anger	Fear	Sadness	Happiness	Disgust
Speech Rate	Decreased for men, increased for female V&K. Increased, Ca, M&A, M, Sch J&L	Increased, Ca, M&A, M, Sch, J&L.	Decreased, V&K, Ca, M&A, M, Sch, J&L (slow).	Increased, Ca, Ma (also decreased), M, Sch, J&L (high).	Decreased, V&K. Ca, M&A

LEGEND	Cahn= **Ca**	Scherer et al= **Sch**	Murray & Arnott= M&A	Ververidis & Kotropoulos = **V & K**	Murray Et al= **M**	Gobl & Ni Chasaide = **G&Ni C**

Table 6: A summary of the findings in the literature reviewed regarding speech rate. There is a strong consensus for all five emotional categories.

3.4 Voice Quality

While a wider definition of prosody that includes all suprasegmental features (not linguistic or verbal) would include voice quality descriptors (sometimes referred to as micro-prosody), voice quality is considered separately here. This is done in order to better define the term *prosody* and limit its use to a few specific acoustically measurable elements, while the term *voice quality* is used to refer to other acoustically measurable variables not as prominent as the main acoustic prosodic parameters. Laver considers these voice quality elements of speech to be a paralinguistic form of communication: aspects of speech that are not part of the spoken language but are required to communicate attitude and emotion (Laver 1994), thus making a distinction between prosody and voice quality all the more important. It is more useful to think of voice quality as part of a layered prosodic system (Vainio 1998).

Like prosody, the term voice quality is used to refer to elements of the speech signal. Whilst not as immediately obvious in speech as the main prosodic parameters, they are associated with the biomechanical mechanisms of speech production (Laver 1994). These micro perturbations arise from the fact that the vocal folds and muscles are unable to sustain a constant period of oscillation. Various physiological process and states affect vocal tract configurations, altering the quality of the voice in numerous ways (2.2.2) with two of the most common perturbations being jitter (micro-perturbations in F0) and shimmer (micro-perturbations in intensity) (Laver

1994; Biersack 2005). Numerous terms are used to describe voice quality: rough, harsh, bright, coarse, creaky, breathy, chest tone, tense etc (Scherer 1986b; Cahn 1990; Murray 1993; Gobl and Ní Chasaide 2000; Gobl and Ní Chasaide 2003; Schröder 2004b) with little or no consensus as to the exact acoustic realisation of each subjective term. However, both jitter and shimmer have been found to relate to perceived "roughness" and "hoarseness" in speech (Laver 1994; Dejonckere, Remacle et al. 1996)[13]. Increased high frequency energy has been found to correlate to voice quality descriptors such as 'sharp' (Juslin and Laukka 2003) and 'harsh' and 'bright' (Breazeal 2004).

	Anger	**Fear**	**Sadness**	**Happiness**	**Disgust**
Voice Quality	Increased brilliance, Ca. Decreased breathiness, Ca. Breathy, M&A. Tense, harsh, G&NiC. Tense voice, Sch. Laryngealisation, M.	Increased brilliance, Ca Irregular voicing, M&A Slight whispery and breathy, G&NiC. Tense voice, Sch.	Decreased brilliance & increased breathiness, Ca. Resonant, M&A. Lax-creaky G&NiC. lax voice, Sch. Laryngealisation, M.	Slight decreased brilliance, Ca. Breathy & blaring, M&A. Tense voice, G&NiC.	Slight increase in brilliance, Ca. Grumbled/chest tone, M&A.

LEGEND	Cahn= **Ca**	Scherer et al= **Sch**	Murray & Arnott= M&A	Ververidis & Kotropoulos = **V & K**	Murray Et al= **M**	Gobl & Ni Chasaide = **G&Ni C**

Table 7: A summary of the findings in the literature reviewed regarding voice quality. There is no apparent consensus for any of the five emotional categories. This may be due to the use of different voice quality descriptors and definitions.

3.4.1 Voice Quality As An Indication Of Emotional State

Research into voice quality in relation to emotion in speech and depression suggests that voice quality does have a role in signalling affect (Johnstone and Scherer 1999; Gobl, Bennett et al. 2002; Ozdas 2004). However, defining a conclusive relationship between voice quality and emotional states is difficult. Voice quality is usually considered alongside the more common prosodic acoustic parameters but Gobl et al. examined voice quality in relation to affective states, finding that no individual voice quality was associated with any one of the states examined (Gobl, Bennett et al. 2002). None of the four specific attributes examined were easily perceived using the

[13] As referenced in (Brockmann, Storch et al. 2008)

voice quality parameters used, but voice quality appeared to be useful for the perception of attitudes and moods rather than emotions. The authors contend that this may have been due to emotions needing more extreme voice quality settings, possibly in conjunction with other prosodic modifications, with research in this area being hampered by the lack of concise definitions for voice quality descriptions. A further study by Gobl and Ni Chasaide reinforced their earlier findings, also concluding there is no one-to-one mapping between voice quality and affect (Gobl and Ní Chasaide 2003). However, different voice qualities did seem to induce different colourings in neutral utterances and the voice qualities studied were better at signalling milder effective states than stronger states. Overall they conclude that voice quality does appear to have a role in judging a speakers emotional state, but it seems that it needs to be considered in conjunction with a wider set of acoustic parameters.

Other researchers have included voice quality along with other prosodic parameters. Cahn found anger to have increased brilliance and decreased breathiness (Cahn 1990). Sadness in contrast had decreased brilliance and increased breathiness. Disgust and fear had increased brilliance and glad speech had decreased brilliance. Murray and Arnott found anger to have a breathy voice quality with a chest tone; happiness was also breathy but with a blaring voice quality (Murray 1993). Disgust also had a chest tone but with a grumbled quality, sadness had a resonant quality and fear was characterised by irregular voicing. Murray and Arnott suggest that voice quality may be important in differentiating between secondary emotions (2.4). Scherer suggests that voice quality is the key to the differentiation of discrete emotional categories and certain voice quality types are associated with different emotions: tense voice is associated with anger, joy, disgust, contempt, grief, anxiety, irritation and fear while a lax voice is associated with sadness (Scherer 1986b; Gobl and Ní Chasaide 2003). Other research has suggested a similar correlation: tense and harsh voices are associated with an aggressive state, while lax, breathy, whispery and lax-creaky voices are associated with a calmer, non-aggressive state (Gobl and Ní Chasaide 2000; Schröder 2004b)

These findings highlight the main problem in comparing voice quality findings: different studies focus on, and describe, different voice quality measurements using a variety of terms. There is no concrete definition for the terms used in describing voice quality and little consensus on the relationship between descriptors and measurable acoustic parameters (Scherer 1986b). It would seem that the use of largely subjective descriptive terms in relation to subjectively termed emotional categories is a major obstacle in determining precise acoustic parameters relating to voice quality (Tatham and Morton 2004).

3.5 Stress

Stress is an oft-discussed aspect of prosody, particularly from a linguistic point of view. From the linguistic perspective, stress is used to distinguish between an emphasised phrase in a sentence or an individual stressed word. However it is not currently possible to accurately determine the exact acoustic correlates of stress in a speech signal, despite the linguistic representation of prosody considering a number of acoustic parameters to be elements of stress (Dutoit 1997) (3.1). This is further complicated by the use of terms such as prominence and accent as synonyms for stress. Laver makes the distinction between phonetic stress and phonological stress: phonetic stress is an indication of the most prominent syllable in a given utterance, while phonological stress, when present on a syllable or word, defines the word-stress for that word (Laver 1994). Laver proposes using the term accent for a normally stressed syllable in a word and stress as the actual stress location. Dutoit uses the two terms interchangeably, while Werner and Keller use the term accentuation as an alternative name for stress (Werner and Keller 1994). Acoustic measurements of stress depend on a number of combining elements: duration, pitch and intensity. The speech analysis software, LinguaTag, uses these three parameters to define stress (5.2.2)(Cullen, Vaughan et al. 2008a; Cullen, Vaughan et al. 2008b). However, the examination of stress in a language is only relevant if that language has stress; while all languages have F0, not all have stress. French is language without stress; it does not use any of the mentioned acoustic parameters to signal phonological contrast, and French late learners of Spanish have difficulty perceiving Spanish language stress (Dupoux, Sebastia´n-Galle´s et al. 2008).

3.6 Discussion

Some researchers have found that emotional states may be manifest in observable physiological states (2.2.2). Focusing on the acoustic realisation of prosody provides defined, measurable acoustic properties that can be used to analyse a speech signal for acoustic parameters in relation to certain emotional states. The argument is made that it is important to separate voice quality from prosody: while it may be part of a broader definition of prosody (Vainio 1998), this research considers voice quality as being separate from prosody in order to provide a more rigid definition of prosody. This allows the term prosody to be used in reference to a number of acoustically measurable variables rather than as a catch all term for all acoustic speech parameters. Some researchers have examined the relation of voice quality to emotional states, finding that voice quality appears to have a role in signalling the presence of an emotional state (3.4.1) (Johnstone and Scherer 1999; Gobl, Bennett et al. 2002; Ozdas 2004). The findings of the various researchers have been combined into a table (Table 8); the emotional categories where there is a strong consensus are highlighted in green: the yellow highlighted text illustrates a weaker level of consensus.

	Anger	Fear	Sadness	Happiness	Disgust
Pitch mean	Increased /high V&K, Ca, M & A, Sch, J&L	Increased/high, V&K, Ca, M&A, Sch, J&L	Decreased/low, V&K, Ca, M&A, Sch, J&L	Increased/high, V&K, Ca, M&A, Sch, J&L	Decreased, V&K, Ca (or equal to), M&A,
Pitch range	Increased V&K, Ca, M&A, M, Sch	Increased, V&K, Ca, M&A,Sch	Decreased, V&K, Ca, M&A, Sch	Increased, V&K, Ca, M&A, Sch	Increased for men, decreased for females, V&K. Decreased, Ca. Increased, M&A
Pitch contour	Rising,Ca, J&L. Raised, M. Downward directed, Sch.	Rising, Ca, J&L. Raised, M, Sch	Rising, V&K. Lowered, M. Downward, Sch Falling, J&L	Falling, V&K. Rising, Ca. Raised, M. Rising J&L.	
Intensity mean	Increased, V&K, Ca, M&A, Sch, J&L	Increased or no change, V&K. Low, J&L (except in panic fear). Increased, Ca, Sch. Normal, M&A	Decreased, V&K, Ca, M&A, Sch, J&L	Increased, V&K, Ca, M&A, Sch, J&L (medium to high)	Decreased, V&K. No change, Ca. Increased, Sch.
Intensity range	Increased, V&K, Sch.	Increased, Sch.	Decreased, V&K, Sch.	Increased, V&K. Increased (Joy) Decrease/no change (Happy), Sch	
Speech Rate	Decreased for men, increased for female V&K. Increased, Ca, M&A, M, Sch J&L	Increased, Ca, M&A, M, Sch, J&L.	Decreased, V&K, Ca, M&A, M, Sch, J&L (slow).	Increased, Ca, Ma (also decreased), M, Sch, J&L (high).	Decreased, V&K. Ca, M&A
High Freq. Energy	Increased, Sch High, J&L.	Increased, Sch. Decreased ('little'), J&L.	Decreased, Sch, J&L ('little').	Medium (slight increase) J&L.	
Articulation Rate	Increased, Ca, Sch. Tense, M&A,	No change, Ca. Precise, M&A. Increased, Sch.	Decreased ("slurring"), Ca, M&A, Sch	Precise with slight decrease, Ca. Normal, M&A.	No change, M&A.
Voice Quality	Increased brilliance, Ca. Decreased breathiness, Ca. Breathy, M&A. Tense, harsh, G&NiC. Tense voice, Sch. Laryngealisation, M.	Increased brilliance, Ca Irregular voicing, M&A Slight whispery and breathy, G&NiC. Tense voice, Sch.	Decreased brilliance & increased breathiness, Ca. Resonant, M&A. Lax-creaky G&NiC. lax voice, Sch. Laryngealisation, M.	Slight decreased brilliance, Ca. Breathy & blaring, M&A. Tense voice, G&NiC.	Slight increase in brilliance, Ca. Grumbled/chest tone, M&A.
Pitch variability	High ('much'), J&L.	Increased, Sch. Decreased, J&L.	Decreased, Sch, J&L ('little')	Increased, Sch, J&L ('much')	
Duration	Decreased, V&K	Decreased, V&K	Increased, V&K	Decreased, V&K	

LEGEND	Cahn= **Ca**	Scherer et al= **Sch**	Murray & Arnott= M&A	Ververidis & Kotropoulos = **V & K**	Murray Et al= **M**	Gobl & Ni Chasaide = **G&Ni C**

Table 8: Summary of the findings regarding the acoustic correlates of the five main emotion categories and the relevant researchers.

There is a strong consensus for a number of acoustic parameters and emotional dimensions (highlighted in green).

As the table illustrates, there is a consensus within the literature regarding a core group of acoustic parameters in relation to a set of five primary emotional states. The strongest consensus is for pitch mean, pitch range and speech rate: all parameters increased for anger, fear and happiness and decreased for sadness. There was a weaker consensus for these variables in relation to disgust, with some studies finding they decreased while others found they increased. For intensity mean there was strong agreement that it increased for anger, fear and happiness, while it decreased for sadness. Overall there is a weaker consensus in relation to intensity range for all emotions, due to the fact that not all studies reported results for it in relation to each of the five emotions.

It must be noted that not all the literature reviewed examined the same acoustic parameters or the exact same emotional categories, though the five primary emotions of anger, fear, happiness, sadness and disgust were included in the majority of them. The lack of a commonly agreed set of emotional categories coupled with different research methodologies potentially confounds the issue: are differences in results due to the different methods used to obtain emotional speech? Or are they due to differences in the understanding and definition of the subjective emotional categories being studied? It is important to note that there is also a disparity between the sources of the emotional speech studied in the literature: some researchers used actors and others broadcast sources, while others opted for Mood Induction Procedures. Mood Induction Procedures (MIPs) are designed to induce emotional states in a subject in a controlled environment (see 4.2 for further discussion on this). The use of actors and broadcast sources may call into question the naturalness and veracity of the emotional speech being examined (see chapter 4).

The findings detailed in the preceding sections support the hypothesis of some researchers regarding a correlation between certain acoustic parameters (pitch mean, pitch range, pitch contour, intensity and speech rate/tempo) and the activation dimension/aspect of emotional states (Murray 1993; Scherer, Johnstone et al. 2003; Schröder 2004b; Mauss, I.B. et al. 2009) (see 2.2.2 for an examination and discussion of the arousal/activation aspects of emotion and 2.5 for an examination and discussion on emotional dimensional models). This relationship appears to be positive: higher measurements equate to a higher level of activation in the subject. Pereira and

Schröder used primary emotional states and circumplex models, showing a relationship between the primary emotional states and certain locations on the models; both found the level of emotional activation to be correlated with the activation dimension of the circumplex models used in their research. Pereira found anger to be situated on the positive end of the arousal dimension and sadness on the negative end of the dimension and determined that pitch mean, pitch range and intensity mean were positively correlated with the level of emotional arousal. (Pereira 2000). Similarly, Schröder, Cowie et al. found mean pitch, pitch range, mean intensity, speech rate and high frequency energy to be positively correlated with the activation dimension of their dimensional model (Schröder, Cowie et al. 2001; Schröder 2004a). A correlation of acoustic parameters with the evaluation dimension was not as strong, though positive evaluation appeared to correspond to a faster speaker rate, less high frequency energy, a low pitch mean and a large pitch range. Schröder, Cowie et al. also discuss the positioning of emotional states within a dimensional space, determining that the acoustic correlates of anger corresponded to the acoustic correlates of its location within the model (Schröder, Cowie et al. 2001). While certain parameters are correlated with the activation aspect of emotion, correlating acoustic parameters with the evaluation aspect of emotional experience has proven more difficult. There also appears to be a positive relationship between pitch, intensity and speech rate; in most cases when one of the parameters is increased so are the other two. While voice quality is important in emotional speech research, there is not a strong consensus in the literature regarding the descriptive terms used and their exact acoustic realisation. This makes investigation of the role of voice quality in relation to emotional states more difficult.

3.6.1 Subjective Categorisation

In most cases, the studies into the acoustic parameters of emotional speech use high-level emotion terms (sad, angry, happy etc) that are considered primary emotions (2.3 and 2.4) This is possibly due to a tacit acceptance of emotional categories which, though not rigorously defined, are used in the main throughout the literature. Many studies are concerned with emotional speech synthesis: creating artificially emotional voices that are judged by listeners to be one of a number of given emotions. Therefore the focus of the studies is in synthesising voices that will sound like naturally occurring speech, which is often described using every day lay terms such as anger,

happiness etc. and other subjective emotional descriptors (2.4). The subjective nature of emotion terms/categories and the lack of a concise definition or even a common consensus regarding their meaning is a big obstacle in studying emotion in speech. How is one to know if two separate authors/researchers mean the same thing when they use the word ‘anger’ to describe an emotional state? While this research is concerned with emotional dimensions (2.5) rather than emotion terms per se, the findings regarding the acoustic parameters of abstract emotional categories are still relevant. The use of these terms within this chapter is not an acceptance of these terms; rather it is done in keeping with the terminology used in the associated literature.

While the words and categories used may be subjective and not always consensual across the various studies, certain acoustic parameters have been used to discriminate between separate emotional categories or abstract emotional classes, regardless of their label. Whether there is a measurable acoustic difference between ‘anger’ and ‘sadness’ or ‘disgust’ and ‘fear’ is important for the fact that there is a measurable difference regardless of the labels applied. In this section, use of the labels ‘fear’, ‘anger’, ‘sadness’, ‘disgust’ etc says nothing about the meaning or exact nature of these labels and makes no claims as to the content of these labelled categories. These labels could arguably be almost anything e.g. w, x, y and z and it would do nothing to diminish the fact that there are measurable acoustic differences between them. Similarly, the focus on a core set of acoustic parameters is due to the fact that they are utilised in the majority of the literature reviewed; while other parameters are also examined, a core group of acoustic parameters is included and examined in almost all the studies. The emotional categories examined in the literature are often considered primary emotions: strong emotional displays that are included in numerous lists of primary emotions and are often placed at the edges of circumplex models due to their extreme nature (2.3, 2.4 and 2.5). While these are important emotional states, they do not constitute the majority of spoken human interaction: the more subtle underlying emotional states are considered a more important and prevalent aspect of human interaction (2.5). While the focus of the majority of studies regarding the acoustic correlates of emotional speech are on these extreme/primary emotional states, it may be more relevant to examine the acoustic parameters of underlying emotional states.

The fact that there is a consensus with regard to certain acoustic parameters and emotional categories is justification for their examination in relation to natural underlying emotional speech.

3.7 Conclusion

The purpose of this chapter was to define the acoustic representation of prosody and to examine the literature relating to the acoustic correlates of emotional speech. This was done in order to inform the analysis of MIP obtained assets. The three representations of prosody were first examined (3.1), detailing the differences between them and focusing on the acoustic representation. A review of the literature regarding the acoustic parameters of certain emotional states was then examined. The review focused on five emotions that featured in the majority of the literature reviewed, along with the most examined set of parameters in relation to those five emotional categories. Voice quality was next discussed (0), arguing that it should be separated from the acoustic representation of prosody, due to the difficulty in ascertaining the precise acoustic parameters related to it. This allows the term prosody to refer to a set of defined, acoustically measurable parameters. A review of the literature regarding the role of voice quality in emotional speech followed, determining that a lack of definition regarding voice quality was an obstacle to determining the voice quality parameters of emotional speech (3.4.1). A discussion on the findings illustrated a consensus between studies regarding some acoustic parameters in relation to certain emotional categories (3.6). Furthermore, a positive correlation was observed between particular acoustic parameters and the activation/arousal aspect of emotional states. While full-blown/primary emotional states are the main focus regarding the acoustic correlates of emotional speech, the examination of the acoustic correlates of underlying emotional states may be of greater relevance due to their argued importance in human communication (3.6.1). The review and discussion carried out in this chapter has given rise to the following research questions:

RQ 2: Can certain acoustic parameters of natural underlying emotional speech be correlated with the activation dimension of a two-dimensional circumplex model?

RQ 3: Can certain acoustic parameters of natural underlying emotional speech be correlated with the evaluation dimension of a two-dimensional circumplex model?

4. Existing Emotional Speech Corpora

This chapter examines existing speech corpora, particularly in terms of sound quality and the source of their assets. Within this chapter and following chapters the term assets is used to refer to segments of audio that form part of a speech corpus and that can be utilised in a number of ways. The term is used widely in professional digital media creation industries and is used to refer to a graphic, animation, piece of video and a sound/piece of music to be utilised for a definite end. The speech segments/clips in the speech corpus developed in this research are assets to be used for a definite end i.e. the examination of acoustic parameters in relation to underlying emotional states.

The various methods and sources used to acquire the emotional assets found in existing corpora are considered in order to inform the best direction to take in obtaining natural underlying emotional speech assets. Assets in existing speech corpora are obtained from simulated, broadcast and elicited sources (4.1). A few examples of claimed 'natural' emotional speech databases exist (Scherer and Ceschi 1997; Chung 1999; Douglas-Cowie, Campbell et al. 2003) although the justification for such content is that it is more 'natural' when compared with simulated content. In the majority of cases what is termed 'natural' emotional assets are obtained from a broadcast source (mainly television) (Douglas-Cowie, Campbell et al. 2003), which calls into question the naturalness of the emotions being expressed. The audio quality of broadcasts varies greatly depending on the nature of the programme, the location and the equipment used (4.1.2); simulated assets often offer a higher level of control over audio quality (4.1.1). Induced assets are obtained using Mood Induction Procedures (MIPs) but are not as widely used as simulated or broadcast assets, but like simulated sources, they offer a greater degree of control over audio quality (4.2). A high quality natural emotional speech corpus is dependent on obtaining authentic, naturalistic, high quality speech assets. Therefore it is important to investigate the nature of simulated, broadcast, and induced assets in relation to the acquisition of naturalistic emotional assets. The use of simulated and broadcast assets will be first considered, followed by analysis of the use of MIPs in obtaining induced emotional assets, arguing that this is the ideal method for

obtaining naturalistic emotional responses from participants (4.2). The use of MIPs also allows a high level of audio quality to be achieved as they take place in a controlled setting and thus can ensure that audio is recorded at as high a quality as possible. The importance of using a high quality level of recording is discussed (4.3), arguing for the use of HD audio standards when possible for the purpose of future proofing and archiving audio recordings.

4.1 Sources Of Emotional Speech

This chapter categorises the sources of emotional speech as simulated assets, broadcast assets or induced/natural spontaneous asset. While there is some crossover between the simulated and broadcast categories, a broad distinction can be drawn between the two. Examining the lists of existing emotional corpora from Ververidis and Kotopoulos (Ververidis and Kotropoulos 2006), Douglas-Cowie (Douglas-Cowie, Campbell et al. 2003) and Juslin and Laukka (Juslin and Laukka 2003) reveals that the majority of emotional speech is derived from non-natural simulated sources (actors, read texts etc) with very little data being truly natural or spontaneous.

Speech Source	Simulated	Elicited	Natural	Synthesis	Manipulated	Semi-natural	Researcher
	8	X	5	X	X	8	Douglas-Cowie
	35	8	17	X	X	1	Ververidis & Kotopoulos
	90	7	12	21	13	X	Juslin & Laukka

Table 9: Summary of tables detailing the source of the emotional data used in a wide variety of emotional corpora. The majority use simulated emotion with only a few using elicited emotion.

Douglas-Cowie et al. list the sources of the data as being either simulated, natural or semi-natural while Ververidis lists an additional source (elicited); Juslin and Laukka also include manipulated and synthesis as sources (Douglas-Cowie, Campbell et al. 2003; Juslin and Laukka 2003). For Juslin and Laukka, natural speech came from recordings of aviation accidents; manipulated speech was acted speech that had been manipulated in some way (filtering, masking, reversal etc); elicited speech was speech

obtained using Mood Induction Procedures (MIPs) (4.2 for discussion on MIPs) and synthesis used synthesised voices, sometimes based on acoustic features copied from real emotional portrayals. The majority of corpora and studies examined by all three used simulated sources to obtain their content and only a small number were described as natural: Douglas-Cowie lists five corpora as being natural, Ververidis and Kotopoulos lists 17, and Juslin and Laukka list 12. However, it is argued in the following sections that what are often construed as natural sources and often claimed to be more natural than simulated sources (Chung 1999; Scherer and Ceschi 2000c; Douglas-Cowie, Campbell et al. 2003), are in some cases actually more akin to simulated sources. Furthermore, Ververidis describes eight of the reviewed corpora as being from elicited sources, but considers elicited emotional speech to be neither natural or simulated, and gives the example of non-professional actors imitating professional actors (Ververidis and Kotropoulos 2006). The argument is made in the following sections that induced or elicited sources are often more natural than either simulated or apparent 'natural' sources.

4.1.1 Simulated Assets

Corpora consisting of simulated assets use acted emotional states, read texts and imagined/recalled emotional situations (Banse and Scherer 1996; Enberg 1997; Amir, Ron et al. 2000; Pereira 2000). However it is not actually known how simulated emotion compares to natural emotion (Douglas-Cowie, Cowie et al. 2000). Simulated emotion that involves reading from a text is not a spontaneous expression of emotion: read speech has been found to have characteristics distinct from spontaneous speech (Johns-Lewis 1986). As discussed in chapter 2, section 2.1.4, emotion is considered to be an important factor in maintaining and negotiating social interaction and relationships: simulated assets are often non-interactive (Banse and Scherer 1996; Enberg 1997; Amir, Ron et al. 2000; Pereira 2000), consisting of monologues with little or no interaction from other agents. The neglect of the social dimension of emotional speech means that obtained assets may contain only a limited range of emotions. Furthermore, the participants used in simulating emotional assets may have subjectively different interpretations of the emotions that they are required to simulate. Russell maintains that there is no evidence that acted emotion is the same as spontaneous emotion and when spontaneous emotional displays are used in place of acted emotional displays, there is a decrease in the level of agreement regarding the

specific emotion being displayed (Russell, Bachorowski et al. 2003); thus acted emotional speech may not be a faithful representation of natural speech (Campbell 2000; Russell, Bachorowski et al. 2003). Since no objective emotional scale exists, it is difficult to compare two subjective acted interpretations of an emotional state.

In the Darwinian and Jamesian emotional perspectives, emotion cannot exist without a corresponding physiological state (2.1); this physiological state is an involuntary response to environmental stimuli. It is possible that simulated emotion lacks the necessary corresponding physiological states normally associated with emotions (2.2.2 and 2.2.3); the evolved survival mechanism that emotion serves (Darwinian or social constructivist) is arguably not present in simulated emotional situations. The underlying physiological state of a subject may also be manifested in spoken communication; as Johnstone argues, emotion can induce changes in speech that the speaker cannot control (Johnstone 1996), with these changes possibly reflecting the underlying physiological changes taking place in the speaker (Oudeyer 2003; Breazeal 2004). Actors are able to achieve the change in speech by voluntarily altering it:

"*Thus acoustic analyses of actor portrayed speech might not provide an accurate description of spontaneous affective speech modulation, which is likely to differ both qualitatively and quantitatively,….*" (Johnstone 1996).

It has also been argued that the authenticity of an emotion is dependent on its source and that a simulated emotion may simply resemble emotion without being a proper instance of any emotional state (Pugmire 1994). It can be argued that the source of acted emotion is therefore not the same as the source of spontaneous or natural (non-simulated) emotion. Similarly the use of emotional recall does not generate natural emotional responses (Amir, Ron et al. 2000). Recall of an emotional event is not the same as experiencing that event: the recall of a fearful situation is not the same as actually being in the dangerous situation that was the original cause of the fear. This distinction is physiologically significant as the context within which the emotion takes place is important (Stemmler 1992; Douglas-Cowie, Campbell et al. 2003): the physiological response during recalled instances of fear and actual instances of fear are not the same (Stemmler 2001; Henkel and Hinsz 2004). Even if the physiological response was exactly the same, Schacter and Singer's two-factor hypothesis theorises

that it is the appraisal of physiological states that gives rise to an emotional state and not necessarily the physiological state itself (Schachter 1962) (2.2.3). Henkel and Hinsz argue that the use of emotional recall assumes that a participants recall is eliciting an emotional response: it is possible that they may be reporting or experiencing an emotional state different than the state that was actually experienced at the time of the incident (Henkel and Hinsz 2004). Section 4.2.1 discusses research demonstrating that the retrospective self-reporting of emotional states is often incorrect and over-exaggerated.

Considering the issues discussed it can be argued that the voluntary nature of simulated emotion and the context in which it takes place undermines its authenticity and its suitability as a method of obtaining natural underlying emotional speech assets.

4.1.2 Broadcast Assets

Some corpora use assets obtained from broadcast sources, mainly television (Chung 1999; Scherer and Ceschi 2000c; Douglas-Cowie, Campbell et al. 2003), the justification being that they are more 'natural' compared to simulated assets. Some of the problems associated with the use of simulated assets are also of concern in using broadcast assets, the use of actors being the most pertinent. However, the use of broadcast assets has focused mainly on programmes such as chat shows, interviews and documentaries: these programmes frequently use emotional recall, as is the case with the Reading-Leeds database (Roach, Stibbard et al. 1998; Roach 2000) and the Belfast Naturalistic Database (Douglas-Cowie, Cowie et al. 2000), which has implications for the veracity of the resulting emotional displays (4.1.1).

Using broadcast assets offers researchers a huge amount of material to choose from, but with the burden of having to negotiate the numerous copyright issues associated with using material of this nature (Campbell 2000). Furthermore, the fact that there is such a large body of material to choose from means that the researcher has to make subjective judgements about the emotional states on display in selecting assets for the corpus. The researcher's judgement of the emotions on display will not necessarily be the same as that of another researcher or even the person originally displaying the emotion. It can be argued that any broadcast is a performance, as the speakers are

usually very aware of the recording process taking place. It is recognised in anthropological research that the presence of a camera distorts the reality of the fieldwork situation (Ruby 1980). As Gottdiener states: *"In using video technology the focal point shifts to the presence of equipment and its effects on rapport. Quite obviously the presence of video equipment is a major source of field contamination" (Gottdiener 1979).*

The presence of a researcher and equipment may also cause people to act differently or even feel constrained in what can be said and done (Geer 1957; Gottdiener 1979). It is possible that this distortion and constraint means that televised emotional displays, like simulated emotion, may only be a facsimile of real emotion. The only way to prevent this distortion is to conceal the equipment and covertly record subjects; however this is a highly questionable practice and ethically unsound (Gottdiener 1979). The distorting effect may lessen over time as subjects become used to being recorded (Erickson 1982). This would suggest that it would be more relevant to use clips taken from the middle or towards the end of a televised program as opposed to clips taken from the start. However, there is an inherent perceptual bias to the recording process (Bellman 1977; Ruby 1980) and, unless a researcher has direct knowledge of the creation of the program, there is no way of knowing if a program was edited in chronological order, or if the footage was taken out of context in order to fulfil a certain editorial agenda. This perceptual bias is inherent in the subjective decisions of the cameraman, the director, the producers and the editor (to name only a few of those involved), and it cannot be known how this affects the final outcome of a broadcast piece. While live broadcasts may have less of a perceptual bias, there is still a bias in the editorial goals of the presenter and the production staff. Even before the researcher has made a subjective decision in selecting clips, a large amount of subjective decisions have already been made in making and broadcasting the program.

Assets taken from broadcast sources can be of varying audio quality, as 'broadcast quality' is a term rather than a definition; one cannot assume that assets obtained from broadcast sources are of uniform quality. Audio quality will also vary depending on the nature of the program, whether it is recorded in a studio or outside in public spaces (as many reality television programs are). Various other factors will affect the audio quality: noise from studio audiences; people talking across each other and

environmental noise from outside broadcasts. The equipment used will also affect the sound quality: different broadcast situations may use different recording apparatus (microphones, cameras etc) and methods.

4.1.3 Discussion Of Existing Speech Corpora

The above section considered the emotional audio assets used within existing speech corpora, arguing that there are three main asset types used: simulated assets, broadcast assets and induced assets. While some researchers have used emotional induction to elicit natural emotional responses, this method has not been used as widely as broadcast and simulated assets in the creation of speech corpora. The biggest problem with simulated and broadcast assets is that their emotional authenticity is debatable. The use of read texts in simulated assets means that the speech and emotion is not spontaneous, with read speech having distinct characteristics from that of spontaneous speech (Johns-Lewis 1986). Furthermore, the source of an emotion is an important factor in emotional authenticity, and it can be argued that the source of simulated emotional states is different to the source of natural emotional states (4.1.1).

There is also a tendency to overlook the important social function of emotion (2.1.4), and thus assets may be limited in their emotional range. The physiological states necessary for the creation of corresponding emotional states are arguably not present in simulated emotional situations, and physiological differences between emotional recall and instances of real emotion are evident (4.1.1). There is also no definitive method to compare simulated emotion to natural emotion, and so the naturalness of simulated emotion cannot be assumed. Using broadcast assets raises issues of copyright and perceptual bias (4.1.2) with the distortion of the situation and constraint felt by the presence of recording equipment being acknowledged within anthropological research (Geer 1957; Gottdiener 1979; Ruby 1980). The use of induced emotional sources is discussed in the next section, arguing that they can be used to induce natural emotional responses and hence are a better alternative to simulated or broadcast sources.

4.2 Mood Induction Procedures

In order for assets to be considered natural for the purpose of analysis, they should be derived from non-simulated and non-broadcast sources and ideally be spontaneous in nature, with audio quality being an important consideration (4.3). The induction of natural emotional responses in a laboratory environment, thus ensuring audio quality can be maintained, is achievable through the use of Mood Induction Procedures (MIPs)[14]. MIPs are procedures that are designed to induce specific emotional states in a test subject within a controlled situation. The emotional states are temporary and relatively specific; MIPs are designed to induce as single and pure an emotion as possible. Gerrards-Hesse et al (Gerrards-Hesse, Spies et al. 1994) carried out an extensive review of MIPs , grouping them into five groups:

4.2.1 MIP Group 1

Emotion is freely generated in this procedure using mental techniques such as imagination or even hypnosis. In this MIP the experimenter does not introduce stimuli that leads to a particular emotional state; rather the test subject induces an emotional state upon themselves through hypnosis (Weiss 1987) or imagination/recall (Strack 1985; Goritz 2007). With hypnosis the subject enters a trance and is asked to imagine a situation in the past where they experienced a certain emotion. Similarly, using imagination or recall, subjects are encouraged to re-imagine an emotional event or situation in order to invoke within them the intended emotional state. As argued in 4.1.1, emotional recall may not be the same as actually experiencing the emotional eliciting event first hand. In addition, Robinson and Clore found that a persons self-report of their current emotional state was more valid than reports about past or future emotional experiences. They found that there was a tendency for overestimation of emotional intensity in retrospective recall, due in part to incorrect beliefs regarding the emotional impact of certain situations; they give an example from the literature demonstrating a misreporting of positive emotions in relation to vacations (Robinson and Clore 2002). These findings suggest that emotional recall or imagination may be unreliable as a method of mood induction

[14] The fact that they are termed mood induction procedures is most likely due to the conflation of terms mood and emotion. Nummenma, for example, refers to them as inducing affective reactions (Nummenmaa and Niemi 2004) See 2.6 for related discussion.

4.2.2 MIP Group 2

This MIP relies on external material to induce emotional states, with the subjects also being instructed to 'get into' the mood suggested by the material. The three MIPs in this group are the Velten MIP (Velten 1968; Banos 2003; Goritz 2007), the Film MIP (Baumgardiner 1988; Kirchsteiger, Rigotti et al. 2006; Rottenberg, Ray et al. 2007) and the Music MIP (Cunningham 1990), with the film and music MIPs being the most successful (Gerrards-Hesse, Spies et al. 1994). The Velten MIP uses a number of written first person statements, with each statement relating to a specific emotion; subjects read the statements and are encouraged to 'get into' the emotional state or mood suggested by each statement. For the Film MIP and the Music MIP subjects are usually encouraged, by any means necessary, to get into the mood conveyed by the film or piece of music and to let the overall emotional tone and feel of either stimuli to guide them in this (Juslin and Laukka 2003)[15].

4.2.3 MIP Group 3

This group is very similar to MIP group two, with the film and music MIPs also belonging to this group. Subjects are presented with emotional stimuli, but are not instructed to 'get into' the mood suggested by the material; they are told to let the material or stimuli induce the emotional state. It is assumed that the material alone will induce this state. The gift MIP (Worth 1987; Ruffle 1999) also belongs in this category and assumes that people will be happy or excited to receive a gift, especially if it is unexpected. Ruffles lengthy economics paper presents a strong argument for the fact that gift giving has an emotional effect on the receiver. However, the effect depends on the expectations of the receiver (Ruffle 1999).

4.2.4 MIP Group 4

MIPs in this group use the fact that some situations can create emotional states, usually through the satisfaction or frustration of a subjects needs. By placing subjects in a situation where certain needs are activated, such as the need to succeed at a certain task, frustrating or aiding the subject in the attainment of their need can induce emotional states. The success/failure MIP uses false feedback (positive or negative)

[15] While not discussing music based MIPs per se, Juslin and Laukka discuss how music individuals can be emotionally affected by music and how music can induce emotional states in individuals.

concerning a subject's performance in a test that they believe is testing their cognitive ability (Forgras 1990). Similar experiments have been used to engender a positive emotional state in euthymic[16] bipolar subjects (Farmer, Lam et al. 2006); false feedback regarding how well they were doing in a task successfully induced a positive emotional state. Nummenmaa and Nemi argue that success/failure procedures involve participants in ways that the film and music MIPs do not, and that these procedures can reliably induce emotional states that are differentiated on the positive-negative axis of a dimensional model. Moreover, the goal of success/failure MIP should be ambiguous enough so that participants are unable to estimate how they are doing without feedback being given (Nummenmaa and Niemi 2004). Nummenmaa and Nemi, along with Henkel and Hinsz (Henkel and Hinsz 2004), argue that the emotional state induced may depend on the participants appraisal of the task: failure should result in a negative emotional state, and success in a positive emotional state. The Social Interaction MIP consists of exposing subjects to an arranged social situation, designed to elicit emotional responses, as it is assumed that the behaviours of others will affect the emotional state of the participant(s) (Westermann 1996). Mueller and Curhan examined whether emotional intelligence had an impact on the satisfaction of a counterpart in negotiations: while not strictly examining social interaction as a mood induction procedure, their findings do suggest that social interaction can be used to induce emotional states. This re-enforces the argument that the behaviour or emotional state of those involved in a social interaction will have an effect on others also involved in the interaction (Mueller and Curhan 2006). In addition, Barsade found that people are *"walking mood inductors"* that influence the mood of others, resulting in emotional contagion within groups (Barsade 2002).

4.2.5 MIP Group 5

In this group it is presupposed that emotions arise out of an unspecified physiological state (see chapter 2 for further discussion on this), and that a cognitive appraisal of a situation determines the quality of the emotion (2.1.3). Therefore, by inducing certain physiological states, emotions can be elicited from subjects. The Drug MIP uses drugs to induce certain physiological states (Manucia 1984), while the Facial Expression

[16] Euthymic is defined as *"Mood in the "normal" range, which implies the absence of depressed or elevated mood" (DSM-IV 1994)*

MIP requires subjects to either smile or frown in order to induce a positive or negative emotional state (James D. Laird. 1982). However, as discussed in section 2.1.2, the induction of an emotional state through posed facial expressions cannot be assumed.

4.2.6 The Use of MIPs In The Literature

Numerous researchers have successfully used MIPs to induce emotional states in participants (Gross and Levenson 1995; Iida, Campbell et al. 1998; Fernandez and Picard 2000; Picard, Vyzas et al. 2001; Kehrein 2002; Johnstone, T, Reekum, C.M van et al. 2005). While emotional induction may be ethically dubious in some cases: Tolmitt and Scherer used slides of skin diseases and injured human bodies to induce emotion (Tolkmitt and Scherer 1986), MIPs possibly offer the best solution for obtaining natural emotional assets, specifically a combination of the success/failure MIP and the social interaction MIP (both in MIP group four) (4.2). The success/failure MIP has shown to be effective in inducing emotion; the true purpose of the task (to elicit emotion) can be hidden from participants, thus avoiding demand effects (4.2.7). The hypnosis, imaginary, Velten, film and music MIPs may have the same problems of authenticity as that of simulated emotion (4.1.1). Furthermore it can be argued that films and music are subjective artistic fields, while one participant might find a certain film or piece of music to have a certain emotional quality or tone, another participant might find the complete opposite. Each persons experience of either a film or piece of music is different, resulting from a myriad of factors such as cultural background, socioeconomic situation and personal experiences (Jain 2004). The gift MIP may induce a natural emotional response, but the emotional range is limited to elation (Gerrards-Hesse, Spies et al. 1994) while the drug MIP is ethically dubious and is an impractical method for obtaining assets. The facial MIP again has problems of authenticity as facial expression is only one aspect of emotional communication and it cannot be supposed that a smile will induce the relevant elated emotional state. The outward smile does not always reflect the internal emotional state, so a corresponding reversal of the order (smile=happy emotional state) is not certain to achieve the desired results (James D. Laird. 1982) (see 2.1.2 for further discussion on this).

4.2.7 MIP Based Demand Effects

While some MIPs have been found to be more successful than others (Gerrards-Hesse, Spies et al. 1994), their effectiveness may be overestimated due to what Westermann terms ‘demand effects’ (Westermann 1996). A demand effect occurs when (a) a participant realises or guesses that the purpose of the procedure is to elicit emotional responses and so pretend to be experiencing the desired emotion, or (b) when a participant realises that there is external manipulation taking place. Demand effects are more likely to occur if the participants are instructed to enter into a particular mood (as with MIPs in MIP group 2): any instruction given regarding the required emotional states can cause a demand effect. The success/failure MIP and the social interaction MIP can possibly be used to avoid the creation of demand effects. The true nature of the experiment is not evident and can be further disguised if needed: participants are engaged in a task or social interaction and can be led to believe that the completion of the task is the purpose of the experiment, or that the experiment is for something other than eliciting emotional responses. Introducing a level of ambiguity into the design of an MIP, thus ensuring that participants need feedback to estimate how well they are doing, also serves to disguise the true nature of the experiment (Nummenmaa and Niemi 2004). In addition, the use of a task-based success/failure MIP removes the subjective nature associated with some other MIPs, and allows the researcher to control and manipulate the experiment in greater detail. By frustrating or aiding the subjects in their task, without their knowledge, they can be guided towards natural negative or positive emotional states (Henkel and Hinsz 2004; Nummenmaa and Niemi 2004), without being aware that a certain emotional state is required, thus avoiding the creation of demand effects.

4.2.8 Computer Games As MIPs

Computer game-based MIPs have successfully been used by researchers to elicit and study emotional speech. Johnstone et al. used the Xquest game that places a player in control of a space ship in a galaxy filled with obstacles (mines and enemies) and rewards (crystals) (Johnstone 1996; Johnstone, T., Reekum, C. M. et al. 2005). The simple purpose of the game was to collect crystals in order to proceed to another galaxy, with rewards being offered (extra ships and points) depending on the number of crystals gathered. There was no external manipulation of the game itself: the design of the game was both conducive and obstructive to the obtaining of crystals and to

progression to each new level. 33 participants played 27 games each, resulting in 891 games in total. The computer game procedure was accompanied by the playing of pleasant and unpleasant synthesised sounds. The results led Johnstone to conclude that there was potential for computers games to be used to induce emotional states in participants that affected their speech (Johnstone 1996). Fernandez used a driving simulator coupled with the verbal solving of maths questions during the simulation (Fernandez and Picard 2000). Manipulation of the experiment was achieved by changing the speed of the simulated car and the frequency at which the maths questions were asked. Oertel et al. used a variety of computer games to elicit a level of physiological arousal in participants: Pacman, Magic Balls, Smashing and Solitaire were all used (Oertel 2004). The goal of their study was to see if a game-based task induced a greater state of arousal than a non-game task, and they concluded that game-based tasks could be used to induce emotional responses from participants. Kaiser and Wehrle carried out a pilot study to investigate the use of computer games to elicit emotional responses from participants (Kaiser, Wehrle et al. 1994; Kaiser and Wehrle 2000). They found that computer games offered advantages over other emotion induction procedures: they are interactive, engaging, and can be manipulated and adapted to elicit different emotional responses. Their results were promising and led them to develop the Geneva Appraisal Manipulation Environment (GAME) in order to create experimental computer games for emotional elicitation (Kaiser 1996). As with their previous findings, they found the interactive game was able to induce distinct emotions in participants.

Rizzo argues that emotional computer agents produce a more emotionally engaging experience for users than non-emotional and more traditional computer systems (Rizzo 2001). As far back as 1976, Weizenbaum noted that people could become emotionally involved in relatively simple computer program, with even a short exposure time resulting in *" powerful delusional thinking in quite normal people"*[17]. While Weizenbaums observations and description of the interaction as 'powerful delusional' needs to be considered in the context of the time, it does suggest that computer interaction can provide and elicit an emotional state in participants. Video games, and in particular violent video games, can be seen as a method of experiencing

[17] Cited in (Rizzo 2001) on page 172. The original source is (Weizenbaum. 1976) Pg. 6-7.

different emotions in a safe and virtual setting: this may be especially attractive to adolescent gamers who are often going through the difficult process of constructing and exploring their personal identity (Jansz 2005). The exploration and experience of different emotional states is aided by contemporary game design which usually aims to immerse the player in the environs of the game. Advances in technology make this prospect increasingly possible, especially considering the graphical and auditory abilities of personal computers and dedicated games machines (Xbox 360, Playstation 3, Nintendo Wii etc). Additionally, players are usually unaware how a game will develop (excluding previous experience of the game) but can quickly learn the positive and negative outcomes and thus may possibly develop an attachment and engagement with said outcomes. This attachment can result in positive or negative emotional states (Juul 2003).

4.2.9 Discussion Of MIP Methods

MIPs have been used by numerous researchers to induce emotional responses in participants (4.2). While there are varying types of MIP, a success/failure social interaction MIP would appear to offer distinct advantages: it avoids the problem of demand effects and the numerous issues regarding authenticity relative to other MIPs, while making recourse to the social communicative function of emotion. Of particular note is Roland Kehrein, who carried out experiments using a success/failure MIP (Kehrein 2002). Kehrein used soundproof rooms and a co-operative task (in this case a Lego construction) with one party giving instructions about what was to be built and the other party following the instructions. By manipulating the Lego available and the time allowed, the participants could easily be hindered or aided in the attainment of their goal. The fact that the participants were seated in separate soundproofed rooms allowed the conversational interaction to be recorded as two separate high quality audio channels. This allowed both sides of the conversation to be analysed, including overlaps. Similarly, various researchers have used computer games in order to induce real emotional states in test subjects, finding that computer games were well suited for this purpose (4.2.8).

A combination of the two experimental designs can offer advantages: using computer games as part of a co-operative, task-based MIPs offer a high degree of control, while the use of separate sound proofed rooms enables high quality audio assets to be

obtained. This approach ensures that obtained assets are more natural compared to simulated and broadcast assets, with emotional responses being induced as a result of the experimental conditions and manipulation. The co-operative or adversarial aspect ensures the social aspect of emotional expression is not neglected (2.1.4) and is in itself a form of MIP (4.2, MIP group 4). The resulting emotional assets can be claimed to be natural and spontaneous, arising out of the manipulation of the task and the interaction of the participants with each other, as opposed to voluntary or knowingly coerced attempts to generate emotional states. Considering the discussion of primary/full-blown and underlying emotions (2.3 and 2.4), this should result in the emotional interaction between participants being underlying in nature as opposed to full-blown (2.3). Underlying emotions are integral aspects of human communication and are arguably of greater relevance; moreover, the elicitation of full-blown/primary emotions may not be ethically acceptable (Campbell 2000).

While simulated assets offer more control over audio quality, the audio quality of broadcast assets can vary. Despite the potential for high quality audio recording afforded by using simulated assets, their emotional authenticity is debatable. MIPs have been successfully used to engender natural emotional responses (4.2.6) yet audio quality is not usually given the necessary attention. A game-based (co-operative or adversarial) success/failure social interaction MIP can be used to engender natural emotional states, avoid demand effects, and record the corresponding emotional speech at a high level of quality. While MIPs may be advantageous compared to simulated or broadcast sources of emotion, they do not offer the perfect solution to eliciting naturalistic emotional speech. As discussed (4.2.6), some MIPs have the same issues of authenticity as that of simulated emotion and the level of control that can be achieved is limited. This is an issue that is discussed further in 8.3.1.

Having examined existing corpora and methods for obtaining natural emotional speech assets, the next section examines the use of high quality audio techniques, and equipment for recording natural emotional speech.

4.3 Recording Elicited Emotional Speech

The purpose of this section is to discuss the advantages of using high quality audio recording techniques and equipment for recording natural emotional speech. The audio industry is, as of 2010, adopting higher quality High Definition (HD) audio standards, and this section argues that adopting the same standard in emotional speech recording is important to maintain standards, to future proof against data redundancy and to ensure assets are recorded at an extremely high level of quality. The first part of this section will examine the technical aspects of sample and bit rates, arguing for the use of high sample and bit rates. This is followed by a discussion of existing and emerging audio formats and systems along with the use of professional audio equipment. The practical aspects of using HD audio, the Redbook CD standard and the mp3 standard is then discussed, examining the use of audio on the Internet in the context of applicable standards and technical considerations. The section ends with a discussion examining the main points raised and concluding that HD audio standards should be used when practically possible, with mp3s being a reluctant compromise when using online listening tests.

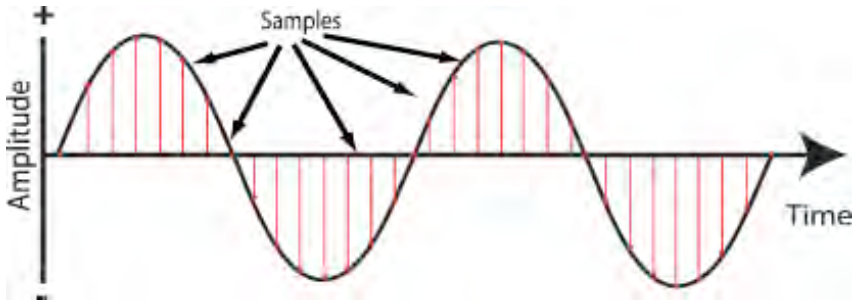
4.3.1 Sample Rate

A digital audio signal is a sampled version/copy of an analogue audio signal. When a sound wave is recorded using a digital system, the system is sampling the analogue waveform to represent it in a binary format using an analogue-to-digital converter (ADC). The converter samples the amplitude of the analogue signal at select intervals, determined by the sample rate (measured in KHz) (Rumsey and McCormick 2002), with higher sample rates meaning that more samples per second are taken. It is important to use an adequately high sampling rate to accurately capture a sound, as failure to do so can result in unwanted audio artefacts, known as aliasing, in the audio signal. The sample rate used should conform to the Nyquist theorem to accurately capture the sound and avoid said aliasing:

"….for any given deformation of the received signal the transmitted frequency range must be increased in direct proportion to the signalling speed, and the effect of the system at any corresponding frequencies must be the same. The conclusion is that the frequency band is directly proportional to the speed" (Nyquist 2002)

Nyquists theorem has become the standard rule-of-thumb when digitising an audio waveform and ideally should be adhered to in any digital recording environment. To accurately sample a sound, the sampling frequency must be at least twice the highest frequency of the sound being sampled (Roads 1996). If too low a sampling rate is used, the digitised waveform will be different to that of the original analogue waveform (Figure 14).

(a)



(b)

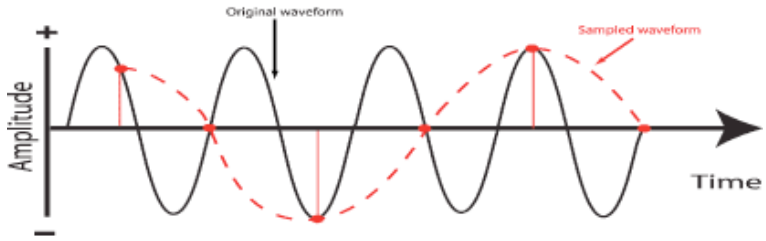


Figure 14: Two examples of a sampled wave. (a) Has a higher sample rate so is a more accurate sample of the original wave. (b) Has a lower sampling rate and the resulting sampled waveform bears little resemblance to the actual waveform, adapted from (Rumsey and McCormick 2002).

As illustrated in Figure 14, a sample rate that is too low results in aliasing. The first waveform (a) is sampled with a sufficiently high sample rate, providing a relatively accurate digitised sample of the original waveform. However, the second waveform (b) has not been sampled with a high enough sample rate and so the sampled wave, when reconstructed and played back using a digital-to-analogue-converter (DAC), is significantly different to the original analogue waveform. The reconstructed waveform bears little resemblance to the original, illustrating that, at the very least, the sample rate needs to accurately capture the cycle of the waveform to avoid a different waveform being digitised (Roads 1996). The more samples that are taken the

closer the digitised waveform gets to the original analogue waveform and the less chance there is that the digitised waveform will contain aliasing errors.

Additional measures can be taken to avoid aliasing; a filter can be added to the signal chain, post-digitisation, to remove frequencies greater than half the sampling frequency; the sampling frequency can also be made slightly higher than twice the highest audio frequency to smooth out the imperfections inherent in the filter (Huber 2005). Higher sample rates also reduce distortion in the audible frequency band introduced into the digital signal by post-production processing: the distortion is spread over the total wider bandwidth, half of which is outside the audible frequency range (Katz 2002).

4.3.2 Bit Rate

As well as a high sample-rate, an adequately high bit rate is also necessary. The standard Redbook CD audio bit rate is 16 bits ((IEC) 1999). This allows for a 16 digit binary number (a 16 digit long string of 1's or 0's) to be used for every amplitude value, resulting in 65,536 possible amplitude values and a dynamic range of 96 dB. Each bit in a system equates to 6 dB of dynamic range and increasing the bit rate allows for even more amplitude values to be represented (along with a corresponding increase in dynamic range): 24 Bits allows for 16,777,216 amplitude values with a 144 dB dynamic range, 32 Bits allows for 4,294,967,296 amplitude values with a dynamic range of 192 dB (Rumsey and McCormick 2002). The higher the bit rate, the more accurately the sampled amplitude values can be represented. If a sampled value lies between two amplitude value steps it is adjusted by the digitising system to the nearest allowable value, a process known as quantisation: the rounded-off value therefore differs from the original value and is referred to as quantisation error or quantisation noise (Roads 1996; Rumsey and McCormick 2002). The amount of quantisation error depends on the bit rate used i.e. the lower the bit-rate gets, the more likely quantisation errors are to occur. Consequently, in order to reduce the amount of quantisation error and noise in the digital signal, a high bit-rate is necessary: the higher the bit rate the lower the error and noise and the wider the dynamic range (Figure 15).

(a)

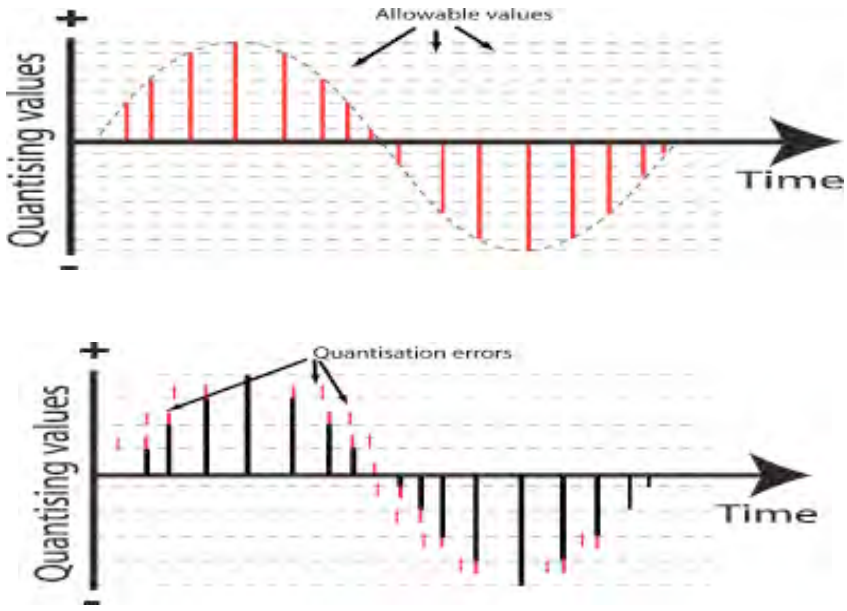


Figure 15: Two examples of a sampled waveform. (a) Has a high bit-rate so the sampled amplitude values are more accurately represented. (b) Uses a low bit-rate, the red bars show the actual amplitude value and the black the quantised value. The sampled amplitude values are less accurate and are rounded off to the nearest allowable value, resulting in quantisation errors/noise, adapted from (Rumsey and McCormick 2002).

A high sample-rate coupled with a high bit-rate is desirable to obtain a detailed and accurate digital representation of the audio signal. In relation to speech research, a high quality audio recording is advantageous as it results in a high fidelity signal with a wide dynamic range: one has to allow for low intensity, low pitched speech (characteristic of sadness) and high pitched, high intensity speech (characteristic of happiness and anger) (3.3), along with non-vocal sounds such as laughter.

4.3.3 Digital Audio Systems And Formats

Various digital audio systems use differing sample and bit rates due to cost and technical and practical considerations. The public switched telephone network (PSTN)[18] operates at 8Khz/8Bit (Titze 1994), the reason being that the PSTN needs to provide a reasonable level of voice intelligibility within a small bandwidth. Since the frequency range of human speech is generally between 125-2000 Hz for males and 300-3400 Hz for females (Talbot-Smith 2001), with human hearing being most

[18] The PSTN is the hierarchical structured worldwide network of the public circuit-switched telephone system, including fixed-line and mobile networks. It connects phone networks in the same way the Internet connects computer networks.

sensitive to a frequency of 4kHz (Kandel 2000), the PSTN sample rate is double the 4kHz sensitive frequency in keeping with the Nyquist theorem. This ensures that the human voice remains intelligible within a small bandwidth, reducing engineering costs and ensuring the network can carry large amounts of phone calls at any one time. Since the range of human hearing is 20Hz-20kHz (Dowling and Harwood 1985; Russ 2008), the Compact disk (CD) Redbook standard specifies a sampling rate of 44.1Khz, ((IEC) 1999). This is more than twice the highest perceptual frequency of human hearing to adhere to the Nyquist theorem and to smooth out the inherent errors of the anti-aliasing filter (4.3.1). The standard Digital Versatile Disc (DVD) specification, both the DVD-ROM audio format and the more common DVD-ROM video format, can use uncompressed Pulse Code Modulation (PCM) audio at various sample and bit rates (from 44.1-192kHz and 16-24 Bit) as well as the compressed audio formats, Dolby Digital (AC-3), and Mpeg 2 audio[19] (Sonic-Solutions 2000). Other audio formats such as DTS and Dolby Digital can be used as long as DVD players support them as they were not part of the original DVD specification, though these are part of the newer Blu-Ray DVD specification (Association 2005). Lossy audio standards are supported to allow flexibility in using the space available on the DVD (the common single side-dual layer DVD has a capacity of 7.95 Gigabytes), striking a balance between the file size of the video files and the audio files, with the DVD Audio format having most of the 7.95 Gigabytes available for audio only.

Similarly, the Super Audio Compact Disc (SACD) format is a high quality audio CD format. It uses an encoding method called Direct Stream Digital (DSD) and operates at a sample rate 64 times that of a conventional CD: 64 x 44.1 kHz= 2.8224 MHz (Spanias, Painter et al. 2007). However, it was not commercially accepted by consumers and remains a niche format. The newer Blu-Ray standard (Association 2005) supports a maximum 192kHz/24bit rate along with various audio standards : LPCM, Dolby Digital, Dolby Digital Plus, Dolby Lossless, DTS surround and DTS-HD (Table 10).

[19] These are lossy audio formats as they compress the audio by removing information from the digital audio file using complex psycho-acoustical based algorithms. Mp3 files are the most widely used lossy format.

CODEC	LPCM	Dolby Digital	Dolby Digital Plus	Dolby Lossless	DTS digital surround	DTS-HD
Max.bitrate	27.648Mbps	640kbps	4.736Mbps	18.64Mbps	1.524Mbps	24.5Mbps
Max.ch	8(48kHz, 96kHz), 6(192kHz)	5.1	7.1	8(48kHz, 96kHz), 6(192kHz)	5.1	8(48kHz, 96kHz), 6(192kHz)
bits/sample	16, 20, 24	16 - 24	16 - 24	16 - 24	16, 20, 24	16 - 24
Sampling frequency	48kHz, 96kHz, 192kHz	48kHz	48kHz	48kHz, 96kHz, 192kHz	48kHz	48kHz, 96kHz, 192kHz

Table 10: This table details the various supported audio formats, bit rates and sample rates of the Blu-Ray DVD standard. Taken from (Blu-Ray-Association 2005).

The increased capacity of the Blu-Ray format allows High Definition (HD) audio and video to be used simultaneously. The use of high sample and bit rates, and the corresponding high quality of the video and audio in the newer media formats means that this level of quality is becoming the standard HD consumer audio standard.

4.4 Professional Audio Hardware

The adoption of High Definition (HD) audio formats is also being adopted by professional audio hardware and software. It must be noted that HD audio is not defined by one particular sample and bit rate. Just as HD video can vary in resolution (from 720p to 1080p) HD audio can range from 88.2 kHz to 192kHz, and 24 bit to 32 bit and beyond. A very widely used hardware and software audio solution, Pro-Tools, can record up to 192 kHz/24 Bit (Digidesign 2009)[20]. The Pro-Tools HD recording system is a professional HD recording system used in numerous professional recording and post-production sound studios. Consumer and semi-professional hardware and software have also adopted HD audio standards. The majority of equipment at this level will not record beyond 96 kHz, but considering the previous discussion (4.3.1, 4.3.2, 4.3.3) this sample rate has offers advantages over the Redbook standard, providing a cheaper alternative to using higher end professional equipment. The very large increases in hard drive capacity coupled with a relative reduction in price makes HD recording systems a more viable professional recording

[20] Other recording software supports high sample rates of 96 kHz/24 Bit and 192 kHz/24 Bit recording as long as the interface hardware supports it. Apples Logic supports 192 kHz/24 Bit, Steinbergs Cubase supports 96 kHz/24 Bit and Steinbergs Nuendo supports 192Khz/24 Bit recording.

option that is ensuring the higher 192 kHz/24 bit level of quality is now the professional HD audio standard.

4.4.1 HD Audio As An Archive Format

With the newer audio formats (DVD Audio, Blu-Ray) becoming the def-facto HD consumer standard, coupled with professional audio hardware and software that operates at this standard, best practice would dictate recording at 192kHz/24bit. While the equipment required to record at such standard is expensive, high sample and bit rates are available using lower cost consumer and semi-professional equipment. As of 2010, the highest sample rate available is 96 kHz (24 bit)[21]. While this is only half of the 192 kHz sample rate, it is still considered HD audio. The argument is made that, when recording a spoken interaction, the highest possible sample and bit rate possible should be used for the following reasons:

- 192-kHz/24 bit HD audio is the professional standard.
- Higher sample and bit rates more accurately capture the original waveform.
- It is always possible to convert down from 192 kHz/24 Bit but it is not possible to convert up to a higher sample and bit rate in the same way[22].
- Higher sample rates reduce distortion caused by digital processing.
- A higher sample and bit rate is ideal for archiving purposes and future proofing current recordings.

4.4.2 Practical Considerations

While high quality audio is desirable, certain practical aspects and applications must be considered. Audio with a 192 KHz/ 24 Bit sample and bit rate takes up a large amount of space. A simple formula can be used to calculate the amount of storage space a digital audio file takes up: sample rate in Hz **x** bit rate in bytes **x** number of channels **x** time (in seconds) (Pohlmann 2000). For example, a one minute stereo audio file at a 44.1 KHz/ 16 Bit level of quality would be: 44100 **x** 2 **x** 2 **x** 60 = 10 584 000 bytes which is 10.09 MB or 10.5 MB per minute (depending on which

[21] The pace of technological advancement means that is likely not to be the case in the very near future.

[22] It is possible to convert an audio file to a higher sample rate. However, no new frequency information is gained and is it is usually only done for compatibility reasons and to ensure playback at the correct pitch and speed.

definition of a megabyte is used: 1000^2 or 1024^2) (IEC 2009); Redbook CD audio uses around 10MB of disk space per minute. Audio recorded at 192 Khz/24Bit results in a file size of between 65-69 MB per minute. Some of the experiments carried out in this research are recorded at this level of quality (7.3 and 8). Each experiment was around ten minutes long, and resulted in a ten minute long dual mono audio recording (in effect a 10 minute stereo file). The file size for each recording is at least 650 MB (325MB per mono channel) compared to a file size of 100MB for a 44.1 KHz / 16 bit quality level. While audio at 192 KHz /24 bit is of better audio quality, it is impractical in certain situations: the large file size precludes its use in any online application and not all audio players can handle the large sample rate and bit rate, thus severely limiting its use. Therefore a more efficient and practical level of quality must be used in certain situations; audio at a 192kHz/24 bit quality level can easily be converted to a lower sample and bit rate.

As noted, the file size of Redbook CD audio is a lot smaller than that of a 192 KHz / 24bit quality audio file. This level is still a widely used standard for many audio files, with the vast majority of audio programs supporting the Redbook standard. The mp3 file format reduces file size even further (depending on the encoding settings, file sizes can range from 1 MB per minute to just over 2MB per minute) and is ideally suited for on-line use. However, one must take into consideration the fact that it is a lossy format[23] and involves the removal of information from the audio file (Hacker 2000). While this is far from ideal, a balance must be struck between the disadvantages of using a lossy audio format and the advantages of using online listening tests, as is the case with this research. This is discussed further in chapter 6, where technical considerations and the advantages of online listening tests are argued to take precedence over the use of a non-lossy audio format (i.e. Redbook Wav).

[23] MP3 compression works by removing data from the audio file, determined using psycho-acoustical, perceptual based codec algorithms: frequencies that are masked by louder and more dominant frequencies are removed before the data is compressed using more traditional methods. See (Hacker 2000)

4.4.3 Discussion Of Audio Quality

High sample and bit rates are desirable for a number of reasons: better audio quality, less post-processing distortion, and HD audio equipment and HD audio formats operate at this standard which is the de-facto professional and consumer HD standard. High quality audio can be converted to a lower sample and bit rate with ease while providing an excellent archive level of quality. While Redbook CD audio was the prevailing audio standard for many years, the current dominant audio standard is harder to discern. While HD audio is standard on DVDs, Blu-Ray, and DVD audio, mp3s are extensively used due to the widespread use of portable music players. Mp3 compression allows for small file sizes with a relatively good level of quality (due to the perceptual encoding algorithms used in the various mp3 codecs) (Hacker 2000). The increased use of docking stations that allow portable mp3 players to be connected to home sound systems, along with stereo systems that can play MP3s stored on data CDs or DVDs, has effectively created two competing audio standards: high quality HD audio and lower quality mp3 audio. As mentioned, mp3 is a lossy audio format, while HD audio is not. This makes the mp3 format an unviable audio archive solution and an unsuitable format for high quality audio recordings. The technical advantages of HD audio compared to mp3 audio are supported in sections 4.3.3 and 4.4.1. While HD audio files use more space compared to Redbook CD quality, the highest sample and bit rate practically achievable should always be used. The decreasing cost and increasing capacity of storage media (Blu-Ray, hard disk drives), coupled with consumer recording equipment adopting higher HD audio standards makes the HD audio format an increasingly viable recording option. While the mp3 format is unsuited for recording and archiving purposes, it is suited for use in online applications (in particular the online listening tests discussed in 6.4 and in 7.3 and 8).

4.5 Conclusion

This chapter examined existing emotional speech corpora, Mood Induction Procedures (MIPs) and current audio formats and standards. It first considered the source of the assets used in the majority of corpora, determining that there are three main sources of emotional speech: simulated, broadcast and induced (4.1.1,4.1.2 and 4.2). While numerous researchers have used emotional induction it is not as widely used as broadcast or simulated sources (4.2). Using simulated and broadcast sources

calls into question the emotional authenticity of the obtained speech. It was argued that simulated sources offer at best a facsimile of authentic emotion. Simulated sources are also problematic in that the source of the emotional response/display is different to that of natural and spontaneous emotional responses, with simulated emotional displays lacking the necessary antecedent physiological conditions. Broadcast sources suffer from some of the same problems as simulated assets and are further problematic as it is argued that any broadcast is in effect a performance and not an authentic behavioural display. The distorting effect of recording equipment has been long recognised in the field of anthropology. Furthermore, the use of broadcast sources is open to perceptual bias and issues of copyright (4.1.2).

Mood Induction Procedures (MIPs) were then examined as a method of emotional induction (4.2). Some MIPs have the same problems of authenticity as simulated sources, and can give rise to demand effects (4.2.1). A success/failure social interaction MIP offers the best option in obtaining authentic emotional speech while avoiding the possibility of said demand effects (4.2.7) as the true nature of the experiment can be hidden from participants (4.2.4). The resulting speech can be claimed to be natural and spontaneous, arising from the manipulation of the task and the interaction of the participants in a controlled social setting. Furthermore, a success/failure social interaction MIP offers the ability to record the resulting speech at a high level of audio quality through the use of sound booths and professional level equipment (4.4). The use of computer games was next examined, showing that a number of researchers have successfully used them to elicit emotional speech from participants and it was concluded that there is merit in using them as part of an MIP based experimental design (4.2.8).

The use of high quality audio was then considered, examining existing and emerging audio standards along with the use of professional recording equipment (4.3). It was argued that high sample and bit rates are desirable, in terms of audio quality, to accurately capture and digitise an audio wave (4.3.1 and 4.3.2). The quality level of 192kHz/24 bit is the professional audio standard and is fast becoming the de-facto standard for the emerging consumer formats (Blu-ray, DVD), and it was argued that the highest level of quality should be used when possible, ideally 192kHz/24 bit (4.3.3 and 4.4.1). This level of quality can be used to archive and future proof the

recordings, while lesser quality versions of the files can be created as and when needed. In some cases lower quality versions are necessary for practical and technical reasons, with mp3s being a compromise between quality and convenience for use in an online listening test (4.4.2).

The review of existing speech corpora carried in this chapter has given rise to the following research question:

RQ 4: Can a practical MIP based experiment be designed to elicit natural underlying emotional speech from participants in a high quality audio environment.

5. Structuring and Annotating an Emotional Speech Corpus

Having examined methods of obtaining natural underlying emotional speech in the previous chapter, consideration is now given to a method to structure, store and utilise emotional speech recordings. MIPs were used to obtain the natural underlying emotional speech assets (chapters 7 and 8) that form the basis of the speech corpus developed within this thesis. This chapter examines methods to fully annotate these assets as they form the basis of the emotional speech corpus. Natural emotional speech assets are only as useful as the data they contain and annotating and documenting this data is vital to make full use of these assets. It is important to understand that the overall corpus creation methodology and its implementation are as important as the speech contained in the corpus. One must be aware that developing emotional speech applications and carrying out in-depth analysis of emotional speech requires a large amount of data that must be structured in a consistent and coherent manner. Unfortunately there is a lack of standardisation for the annotation of speech corpora, especially in relation to emotional content. Coupled with the fact that speech corpora are usually created for a specific purpose, consequently limiting their usefulness and reusability, a coherent annotation schema is vital to enable intelligent and meaningful querying of corpora. This ensures their usefulness and reusability, while enabling wider access and optimised utilisation.

This chapter argues that a three-tiered approach to the annotation of speech corpora, incorporating metadata annotation, acoustic data annotation and emotional data annotation, is necessary to comprehensively annotate emotional speech corpora. Figure 16 gives a conceptual overview of the three-tiered approach to the annotation of Emotional Speech Assets (ESAs).

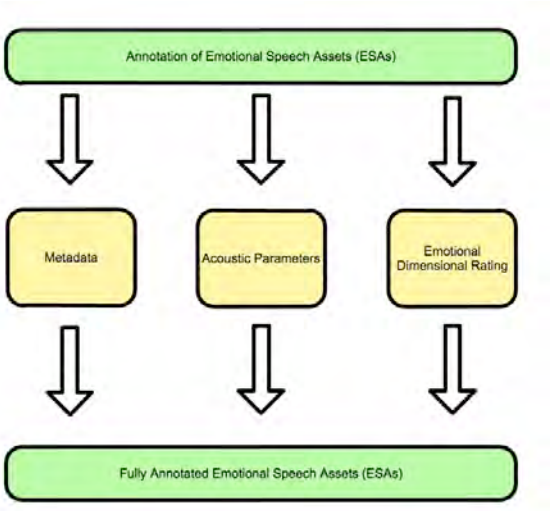


Figure 16: A conceptualisation of the three-tiered approach to the annotation of ESAs proposed in this research. The diagram illustrates the conceptual three-tiered approach in annotating emotional speech assets. Annotating the metadata, the acoustic data and the emotional data results in a fully annotated set of assets.

An overview of existing metadata schemas is first given, discussing their merits and shortcomings (5.1). This is followed by an examination of the Isle Meta-Data Initiative (IMDI) schema and argues that it is particularly suited to the annotation of speech assets obtained using a task based success/failure social interaction MIP. A method of analysing and annotating Emotional Speech Assets (ESAs) for acoustic properties is then discussed, suggesting that the acoustic information can be annotated in a similar manner to the metadata (5.2). This is followed by a discussion of methods for rating and annotating the emotional dimensions/content of ESAs and considers the advantage of using a large sample size to determine the emotional dimensions of collected assets in contrast to the usual method of using a small group of expert listeners (5.3).

5.1 Existing Metadata Annotation Schemas

There are a number of language resource based metadata initiatives in existence: the Dublin Core Metadata Initiative (DCMI) (Core 2009), the Open Language Archive community (OLAC) (OLAC 2008), the ISLE Metadata Initiative (IMDI) (IMDI 2007) and Mpeg7 (MPEG 2004) (Wittenburg, Peters et al. 2002). Two emotional metadata initiatives also exist: the HUMAINE Emotion Annotation and Representation Language (EARL) and the Emotion Markup Language (EmoML) (Baggia, Burkhardt

et al. 2009). The DCMI concentrates on web based resources, providing fifteen elements that can be used to describe authored web resources; contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title and type. There is nothing specific to language or speech within the DCMI schema. In contrast, the OLAC and the IMDI schemas contain elements for speech and language resources. The OLAC is based on the DCMI, building upon it to meet the needs of language resources that are not met by the DCMI standards. This is achieved by adding an extra element to describe the language covered in a resource (OLAC 2008). The MPEG7 Multimedia Content Description Interface is a standard created by the Motion Picture Experts Group to describe multimedia content, allowing the metadata information to be accessed by a device or computer when needed. MPEG7 provides high and low level annotations of the media, allowing description schemes to be defined and covers more than a hundred descriptor schemes (MPEG 2004). It is intended mainly for use in the film and music industry, and it does not contain the elements necessary for specialist speech and language researchers (Wittenburg, Peters et al. 2002).

The Emotion Annotation and Representation Language (EARL) was a proposed syntax for an XML based emotional annotation language (Schröder, Pirker et al. 2006). EARL has not been fully implemented and remains at the proposal stage. Similarly, the Emotion ML schema is an emotion mark-up language and, as of January 2011, is under development and is not yet implementable (Baggia, Burkhardt et al. 2009).

Metadata Schema	**DCMI**	**OLAC**	**MPEG 7**	**IMDI**	**EARL**	**EmoML**
	Concentrates on web based resources as opposed to speech/language resources.	Builds upon the DCMI schema by adding a language element.	Designed mainly for use in the audio visual industry	Describes speech resources in greater detail. Particularly suited to annotating MIP derived speech.	A proposal for emotional corpus annotation. IN the incubation stage.	No fully implement able. In the incubation stage.

Table 11: A brief synopsis of the five metadata annotation schemas examined. The IMDI schema is most suited to the annotation of MIP derived assets. The EARL and EmoML schemas are still in the incubation stages with the EmoML being the most recently developed.

While the various schemas use different approaches, limited mappings between schemas are possible. The IMDI schema describes speech and language resources in greater detail for specialists in that domain, and so is more suited to the annotation of speech assets; moreover, the sub-schemas are also ideally suited to annotating MIP derived speech assets, specifically success/failure-social-interaction MIP derived assets (4.2). The flexibility of the user-defined keys in the schema, particularly in the content sub-schema, allows for the multitude of induction methods and setups possible with MIPs to be recorded. This is an important consideration in examining metadata schemas in relation to corpora composed of such assets: concise details regarding the type and format of the MIP used also need to be annotated. Of the schemas discussed, the IMDI schema presents the best option for achieving this.

5.1.1 The IMDI Schema

The EAGLE/ISLE consortium has developed the ISLE Metadata Initiative in an effort to create a cohesive corpus metadata standard, adopting a different approach to the schemas discussed. The IMDI schema is an extensive schema designed to allow interoperability for multi-media and multi-modal corpora and to provide a means for browsing and searching said corpora (IMDI 2007). A number of existing archives have implemented the schema and can be found at the IMDI Metadata Domain online portal (IMDI 2007) Figure 17 gives an overview of the IMDI schema and the sub-schemas within it.

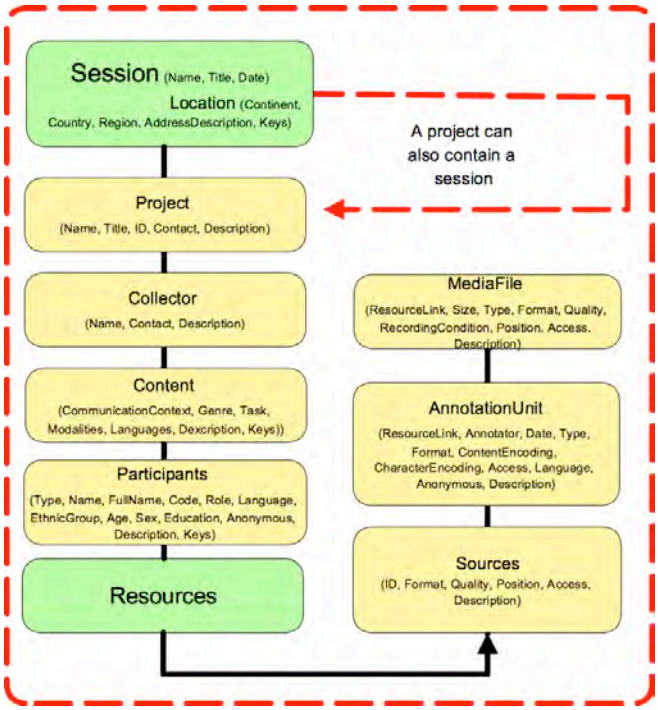


Figure 17: Overview of the IMDI metadata schema and the various sub-schemas within it. The main session element is the parent, containing all the child sub-schemas.

Within the overall session sub-schema are a number of other sub-schemas: location, project, collector, content of the task, participants and details of the resources used (the type of media file and its technical attributes). A session can also be created within the context of a project, and thus a project can contain numerous different sessions

The IMDI schema offers a set of detailed descriptive sub-schemas suited to language and speech resources, with only a few elements being mandatory, providing a degree of flexibility in implementing the schema (Cullen, Vaughan et al. 2008; Cullen, Vaughan et al. 2008c). The IMDI schema is discussed here relative to its use in the creation of an emotional speech corpus, as it is particularly suited to the annotation of MIP derived speech assets. The implementation of a modified IMDI schema is discussed in 6.2, showing how the various elements of the schema were adapted and implemented to annotate the metadata of the ESAs obtained from a task-based social interaction MIP (section 7.3 and chapter 8).

5.2 Annotation of Acoustic Parameters

Annotation of the assets within the speech corpus must also consider the acoustic parameters of the assets. While the IMDI annotation schema allows for a large amount of descriptive data, it does not allow for comprehensive acoustic annotation, certainly not enough for this research (9). The acoustic analysis of ESAs can be carried out using a wide range of software, with numerous packages providing detailed acoustic data. Discussion on this aspect of the acoustic annotation will focus on PRAAT, as it is one of the most flexible and widely used pieces of speech analysis software, and a third party application, LinguaTag, that uses PRAAT as a backend analysis engine.

5.2.1 PRAAT Speech Analysis Software

One of the most widely used software packages within the speech and audio research field is PRAAT (Boersma and Weenink 2006). PRAAT is a cross-platform software application that is capable of analysing, synthesising and manipulating speech, and it can carry out a comprehensive analysis of an audio signal. One of the strengths of PRAAT is that it allows for the creation of scripts to carry out various forms of acoustic analysis and manipulations of an audio file. Nearly all its functionality can be accessed and utilised using these scripts; this enables PRAAT to interface with and be accessed by third party custom applications using these scripts. Furthermore, third party applications can incorporate PRAAT as a backend analysis engine, negating the necessity for end users to have a separate install of PRAAT on their systems.

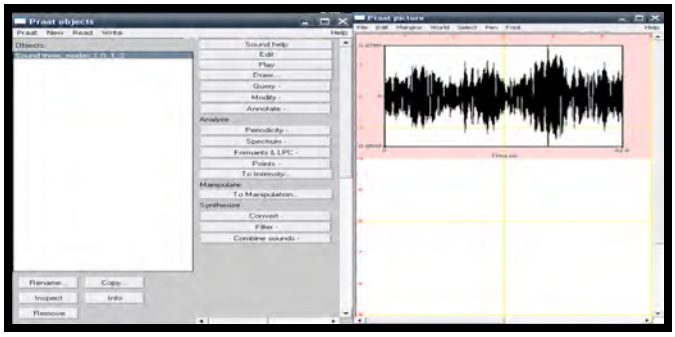


Figure 18: PRAAT screen-shot showing the main interface screen and graphical output screen. The results of most analysis procedures can be drawn in a picture window (right).

PRAAT can provide a wide range of acoustic measurements including: pitch range, pitch contour, mean pitch, intensity range, intensity contour, mean intensity, jitter and shimmer measurements, HNR values, spectrograms, spectral energy distribution. Some of these parameters have been shown to be important in the analysis of the acoustic parameters of emotional speech (chapter 3).

5.2.2 LinguaTag Analysis Software

The PRAAT analyses engine can integrate with the Eiffel object orientated programming language (Software 2008). This has led to the development (locally) of an audio analyses tool named LinguaTag, a piece of software that incorporates the PRAAT analyses engine (Cullen, Vaughan et al. 2008a). LinguaTag displays aspects of the acoustic parameters of an asset in a GUI and isolates vowel events in an audio file; the isolation of vowel events is a common method in speech analysis (F. Pellegrino. 1999; Ramus, Nespor et al. 1999; Honorof and Whalen 2005; Tsao, Weismer et al. 2006) and presents acoustic information about each vowel event. LinguaTag adopts the vowel stress analysis method developed by Cullen et al. (Cullen, Vaughan et al. 2008b): vowel stresses are determined using a combination of duration, pitch and intensity measures, with the user defining the detected stresses as being either primary, secondary or tertiary. The software also allows a speech asset to be annotated for linguistic features and emotional dimensions if needed. The complete acoustic analysis and list of an assets acoustic parameters, along with the dimensional emotion rating and linguistic features, are outputted in the Synchronized Multimedia Integration Language (SMIL 2008) XML format (XML 2008). The SMIL XML format is widely used and is capable of being automatically parsed by numerous backend database technologies (6.3). Figure 19 shows the workflow of the LinguaTag software.

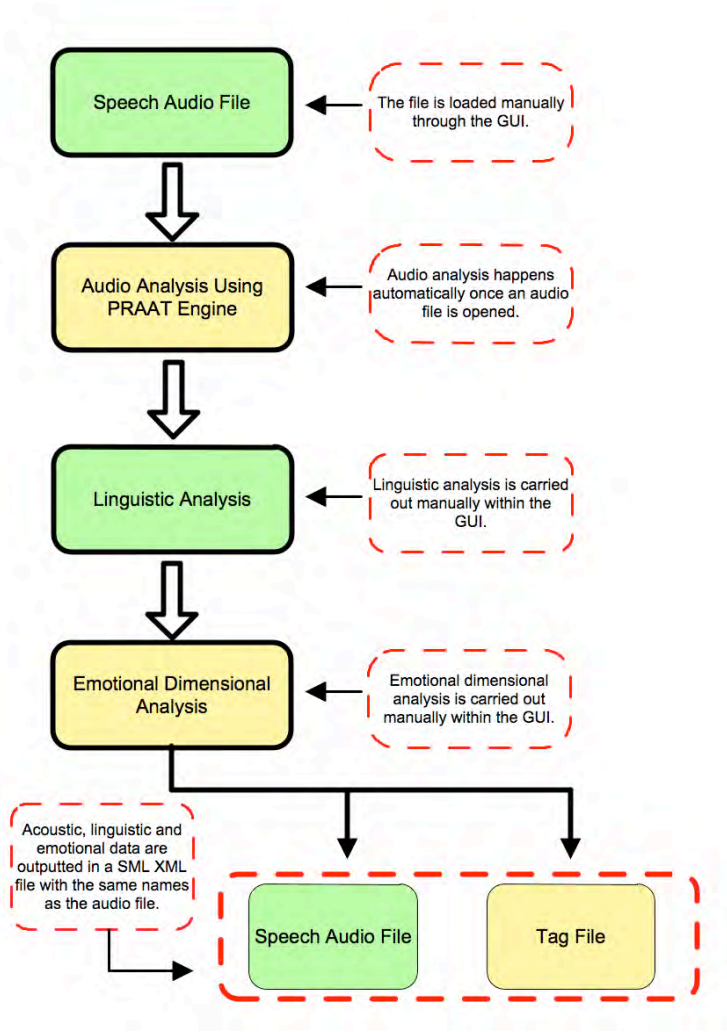


Figure 19: LinguaTag workflow diagram. There are three types of annotation possible using LinguaTag: acoustic, linguistic and emotional. The acoustic annotation occurs automatically one an audio file is opened. The linguistic and emotional annotation is carried out manually. Adapted from (Cullen, Vaughan et al. 2008a)

The outputted SMIL file contains information about the jitter, shimmer, voice breaks, Harmonic-to-Noise Ratio (HNR), intensity and pitch for each vowel event. While LinguaTag allows for a certain level of acoustic analysis, the extent of this analysis is limited. The main significance of the LinguaTag application is its ability to provide acoustic analysis results in a file that can be parsed into a backend database, thus demonstrating a method for the semi-automatic annotation of acoustic parameters.

5.3 Annotation of Emotional Dimensions

Having considered methods for annotating metadata and acoustic annotation schemas, methods of annotating the emotional dimensions of ESAs must be considered. The

emotional content of existing speech corpora is usually decided or rated, in the main, using perceptual listening tests (Enberg 1997; Alter, Rank et al. 2000; Kienast 2000; Pereira 2000). In the majority of cases, the listening groups are relatively small, ranging from four (Polzin 2000) to 12 (Niimi 2001) and 13 (Edgington 1997) to 20 (Enberg 1997; Alter, Rank et al. 2000; Kienast 2000) to 30-plus (Abelin and Allwood 2000; Pereira 2000; Scherer and Ceschi 2000c) to 73 (Wendt 2002). In one case the listening group was as large 1045 people (Iriondo 2000), but this is the exception rather than the norm. Examining some of the corpora in the reviews carried out by Ververidis (Kotropoulos 2003), Douglas-Cowie (Douglas-Cowie, Campbell et al. 2003) and Juslin and Laukka (Juslin and Laukka 2003) shows that most groups tend to be rather small. As Figure 20 illustrates, the majority of listening groups (used in 24 listening experiments) are small, with the average listening group size being 28. Only three of the 24 experiments use relatively large listening groups: the majority use between 10 and 30 people.

Listening groups are not used in all cases. Emotional portrayals by actors are not always subjected to judgement by listening groups, perhaps because it is assumed that the instructions given to an actor are enough to categorise the resultant emotional portrayal. This is based on the assumption that instructions to portray a certain emotion, coupled with an actor's ability, are enough to result in authentic emotional displays. However, as argued previously, the use of acted emotion and the assumption that it results in authentic emotion is problematic and certainly not a given (4.1.) One must also be aware that presenting a set of pre-defined emotional categories to listeners is limiting and constitutes nothing more than a discrimination test or a forced choice listening test rather than a perception test (Russell 1993; Russell 1994; Russell, Bachorowski et al. 2003). As Scherer and Johnstone et al. state: *"…. studies can be criticized for using only a relatively small number of emotions……thus constituting discrimination studies (deciding between alternatives) rather than recognition studies (recognizing a particular category in its own right)"* (Scherer, Johnstone et al. 2003).

Presenting a listener with a limited set of categories may lead to a mislabelling or wrongful categorisation of an ESA: a listener might not agree with any of the pre-defined categories but is forced to use a 'best-fit' in the absence of any other option, thus colouring the results. The use of a set of emotional categories (2.3) in listening

tests is impractical as does not necessarily cover every possible emotion state. Ideally a listener engaged in a perception test should have as few limitations and pre-defined labels as possible imposed upon them. The use of an emotional dimensional model (2.5) is ideal in this regard as it avoids the problems associated with subjective emotional terms and categories. The circumplex dimensional model allows corpus assets to be rated in a simple and straightforward manner: listeners listen to the assets and judge the value on both scales of the circumplex model.

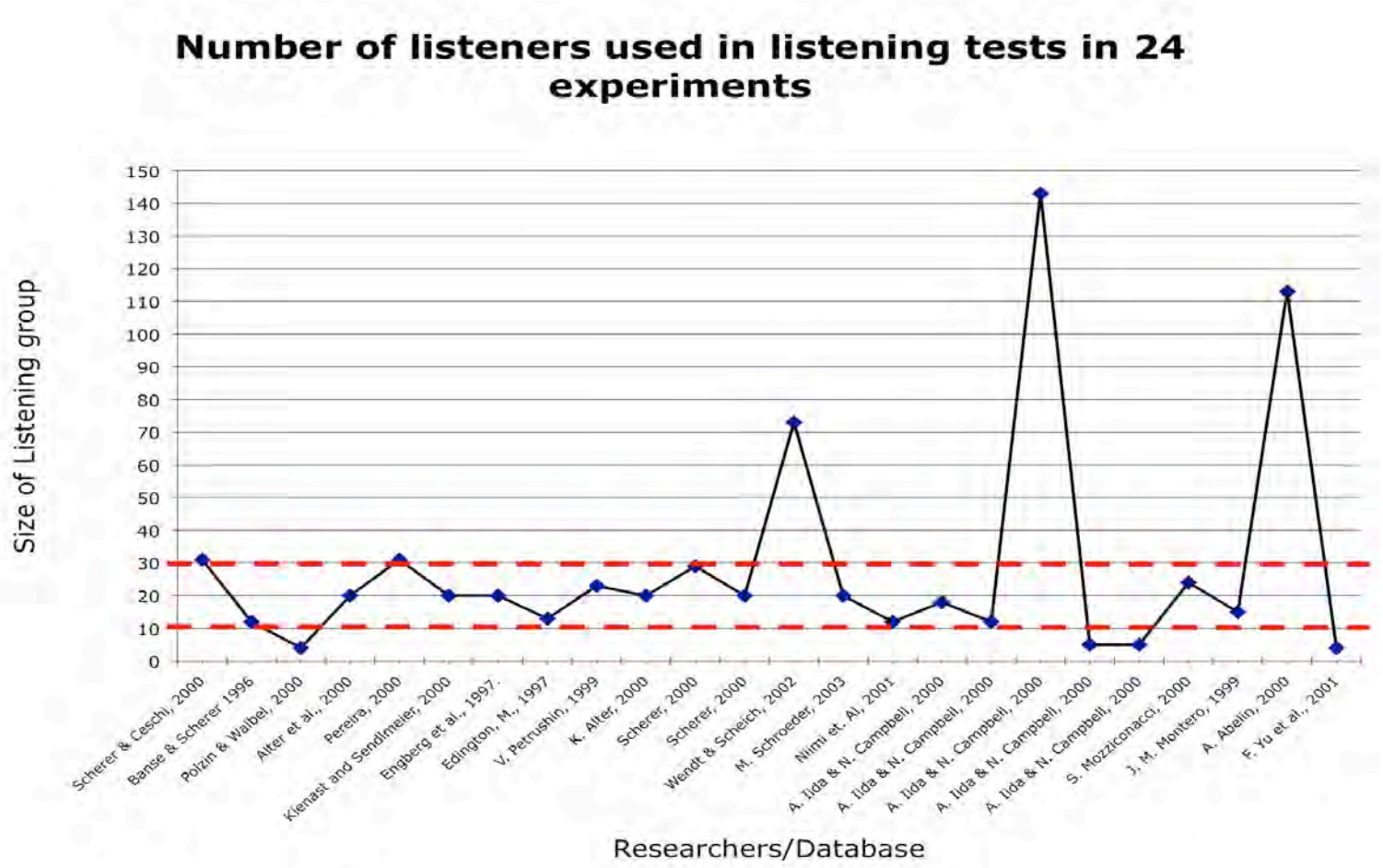


Figure 20: Graph showing the size of listener groups in 24 listening tests. (Iriondo 2000) has been left out. Most listening groups are between 10 and 30 people in size.

Some listening groups employ what are sometimes termed ‘expert listeners’: listeners who are claimed experts in the recognition of emotional states and are usually researchers in the wider field of emotional research. As previously discussed (2.1 and 2.3), emotion is an important aspect of human communication and survival, both evolutionarily and a socially, and it can therefore be argued that each and every one of us is capable of recognising emotional states. Barring cognitive impairment or mental illness, the recognition and understanding of emotion forms a central part of our interaction and engagement with other people (2.3 and 2.4 for further discussion on this topic). The inability to recognise the emotional state of another person is usually associated with some form of psychiatric disorder such as, autism, narcissistic personality disorder or anti-social personality disorder (DSM-IV 1994). For this reason, it is argued that a wider listening group of non-experts can be utilised, providing many distinct advantages over smaller listener groups, expert or otherwise.

5.3.1 Large Sample Sizes As A Method Of Rating The Emotional Dimensions of ESAs

Large listener groups have many advantages over smaller listening groups. Using large sample groups (large listening groups) reduces the probability that a chance relationship between variables will be found (Hill and Lewicki 2005; Peat and Barton 2005). As the sample size increases, the probability of a random deviation from the sample mean decreases along with a decrease in the margin of error (Boslaugh 2008). Using too small a sample size can lead to over fitting: when the sample size is too small there is a danger that the data might fit too well with a given model or hypotheses and thus not be a true representation of the reality of the situation (Boslaugh 2008). In the case of emotional speech perception and rating, results from a small sample size could give rise to an incorrect set of emotional dimensional ratings and therefore would not be a true representation of the emotional dimensions of an ESA. Obtaining a large amount of emotional dimensional ratings results in a more statistically robust and confident set of results whilst avoiding incorrect ratings and the errors inherent in using a small listening group.

5.4 Discussion

The annotation of emotional speech assets (ESAs) should conceivably happen at three different levels: metadata, acoustic data and emotional data. Speech corpora are often created for a specific purpose and as a consequence, their extended usability and relevance may be limited. The comprehensive annotation of corpora arguably addresses these shortcomings as corpora can be utilised beyond their original remit and a comparison between different methodologies is made easier. Large-scale listening tests provide for a more robust set of results and are advantageous to using a small group of listeners. Chapter 6 discusses the implementation and instantiation of the methods discussed in this chapter, illustrating that the approach facilitates the use of web-based interface tools for visualisation purposes and to carry out large scale listening tests.

5.5 Conclusion

This chapter examined methods of annotating speech corpus assets in terms of metadata, acoustic data and emotional data. A lack of standardisation regarding corpus annotation preceded a discussion on existing metadata schemas with the IMDI schema being claimed to be ideal for metadata annotation (5.1.1). The annotation of acoustic data was next considered, with the PRAAT software being discussed as a method of analysing the acoustic parameters of ESA's (5.2.1). The result of LinguaTags limited acoustic analysis procedure is stored in a SMIL standard XML file that can be parsed into a backend database (5.2.2). The advantages of using large scale listening tests was next considered, determining that using a dimensional emotional model in large scale listening tests allows them to be truly perceptual rather than discriminatory (5.3). The implementation and structuring of the three-tiered annotation approach espoused in this chapter is discussed in the next chapter.

The review of annotation methods has given rise to two research questions:

RQ 5: What are the practical considerations of annotating an emotional speech corpus?

RQ 6: What are the advantages and limitations of using a large population size in rating the emotional dimensions of speech assets?

6. Implementing A Three-Tiered Approach To Corpus Annotation

The previous chapter examined methods of annotating speech corpora, arguing for a three-tiered annotation methodology. Methods for the annotation of metadata, acoustic data and emotional data were also examined. Consideration is now given to the technical architecture needed to structure a three-tiered annotation methodology in a logical and coherent manner. This chapter examines the practical implementation of each aspect of the three-tiered approach discussed in chapter 5. The structure of the annotation data is arguably as important as the data itself. Utilising existing database and web technologies, in conjunction with the three-tiered annotation approach, ensures the usefulness and reusability of speech corpora, enables an optimised implementation of the annotation data and enables wider access to the corpus.

The creation of a persistent backend database is first considered, examining current web-based database technologies (6.1). This is followed by a discussion of the implementation of an adapted IMDI schema into the database (6.2). Further discussion of the acoustic annotation of the corpus assets follows, examining LinguaTags SMIL output as a method of incorporating the acoustic annotation into the database (6.3). The creation of a web-based listening tool to carry out large-scale listening tests is then discussed (6.4), along with a prototype interactive corpus visualisation interface (6.4.4). The listening tool and visualisation interface demonstrate methods by which the structured corpus, underpinned by a backend database, can be utilised by web-based applications for various purposes.

This chapter contributes to the answering of research questions: RQ 2, RQ 3, RQ 5 and RQ 6.

6.1 A Persistent Backend Database To Structure A Three-Tiered Annotation Approach

There are several considerations needed to define the technical architecture of the corpus. Firstly, the corpus must allow for the insertion of assets in the form of audio files and the related LinguaTag data (in the form of SMIL files) and mp3 files (for use in online listening tests). The corpus must be able to parse the SMIL files and populate database tables with the information. Secondly, the corpus must also allow the IMDI metadata to be entered and reused as necessary, either for a single Emotional Speech Asset (ESA) or for a batch of ESAs. Finally, the corpus must allow the upload of mp3 audio files (4.4.2) to be used in online listening tests (6.4.2). A corpus, therefore, necessitates a storage layer or database as a persistent back-end, which will allow access to the assets via web-based tools and applications.

One of the most widely used and ubiquitous pieces of database software is MySQL (Sun-Microsystems 2009) which is compatible with a wide range of web programming languages and frameworks, such as PHP (Lerdorf 2006; Group 2009) and Ruby-on-Rails (Hansson 2009). MySQL databases can be searched and queried using standard SQL (Sun-Microsystems 2009) commands; both PHP and Ruby-on-Rails are able to interface with MySQL databases using SQL commands. PHP is used by a large amount of websites, some of which are the largest and most prominent sites on the World Wide Web, and is one of the most popular web programming languages available (Lerdorf 2006). While both languages are suited to the creation of a corpus backend database, Ruby-on-Rails is specifically designed to make common web development tasks easier, allowing complex web applications to be developed and implemented promptly. For this reason Ruby-on-Rails was used in the creation of the backend database. A central requirement for a backend database is to reduce the overhead associated with annotating and uploading assets. Often metadata can be reused for multiple assets, and so should not require the re-entering of this data every time a new asset or batch of assets are uploaded. Ajax provides a way to reduce the overheads associated with entering large amounts of data (Holdener 2008). The Ajax autosuggest facility allows previously entered data to be reused quickly and efficiently, thus greatly increasing productivity and streamlining any data entry procedure (6.1).

6.2 Adapting And Implementing Aspects Of The IMDI Schema

While the IMDI schema is extensive (5.1.1), not all elements need to be implemented. Whereas it makes sense to include certain elements (project, session, collector, content and participant), the inclusion of other elements is not always necessary. The use of only some sub-schemas enables flexible and in-depth metadata to be obtained; other sub-schemas can be added at a future date without affecting existing sub-schemas and data. This allowed the initial implementation to focus only on the sub-schemas that were the most relevant. This also applied to the elements of each sub-schema; some were be used while omitting others. This is one of the main advantages of the IMDI schema; not every sub-schema and corresponding list of elements needs to be used for the annotation of MIP derived assets. Furthermore, unique user defined keys can be used to tailor the schema as required. The implementation of a set of modified sub-schemas allowed corpus assets to be annotated in a detailed manner while also allowing for other sub-schemas and elements of the IMDI schema to be implemented at a future date as necessary.

For this research, four instances of the IMDI schema (project, session, content and actor) were modified and implemented in the database structure and upload procedure. These four sub-schemas were deemed to be the most relevant to the corpus assets at the time of its creation. While the resource sub-schema is used to annotate the technical information about a media file (file-type, quality etc) it is not suited or detailed enough to annotate the acoustic properties of ESAs. This aspect of the schema is not implemented: taking into consideration the acoustic parameters discussed in chapter 3 a more detailed analysis and acoustic annotation method was necessary to annotate the acoustic parameters of each ESA.

6.2.1 Project And Collector

This structure of the IMDI schema is advantageous in that the definition of a particular project allows various sessions and related elements within to be grouped in a consistent form. In the case of an emotional speech corpus, all sessions can be organised relative to the project. Grouping sessions logically allows for future expansion of the speech corpus to include other projects developed for different

purposes. It also allows a specific project to grow in scale by adding more sessions without affecting existing sessions within that project. The project information pertains to the project as a whole and all sessions within that project: similar to an object-orientated language, the classes within the project (session, actor and content) inherit the overall project metadata alongside their own. Elements of the collector element can be implemented at this stage if necessary, as a singular collector can collect all the data for a certain project. Consequently, an adapted project sub-schema was implemented and amalgamated with the collector sub-schema to define metadata relative to a project as a whole. This allowed a single contact source to be specified for each collector, along with descriptive details pertaining to the whole project and every sub-schema within it. Figure 21 illustrates the project and collector sub-schemas and their instantiation in the backend database.

(a) (b)

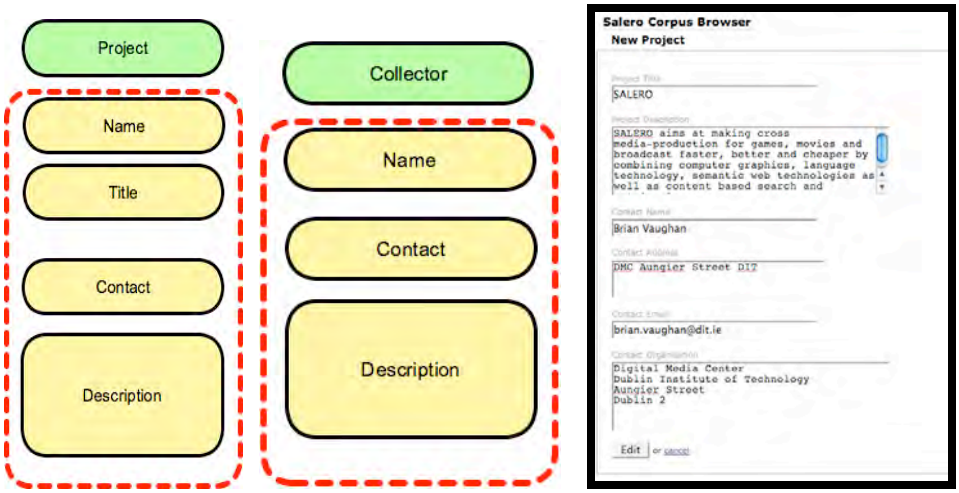


Figure 21: The project and collector schemas as defined in the overall IMDI schema with basic information about the project and the collector (a). The collector and project sub-schemas are implemented in the backend database with only one collector as point of contact, serving as project leader (b).

The collector information relates to the person collecting the data by running a session or a procedure to collect data for the corpus; this could be audio or video and only the name, contact details (email or phone number) and a description of the collector are recorded. In the IMDI schema, a project can have one or many collectors and one collector can collect all the data for a project. Essentially, the two different sub-schemas were amalgamated due to their elements being similar and to ensure that

each project specified only one person as a collector and a point of contact. This allows one person to act as a project leader to coordinate all aspects of the project but does not necessarily preclude other researchers from collecting data.

6.2.2 Session

A session is described as the common bundle for speech events with all speech assets being defined relative to a particular session. This allows a longer clip to be broken down into assets within the same session and retain the session metadata; individual assets to be examined in isolation or in the wider context of the antecedent session. The session definition is an expedient way to group assets for analysis, allowing assets taken from the same or different sessions to be assessed, either in isolation or within a wider related context (e.g. along project lines). Figure 22 illustrates the session sub-schema and its instantiation in the backend database.

(a) (b)

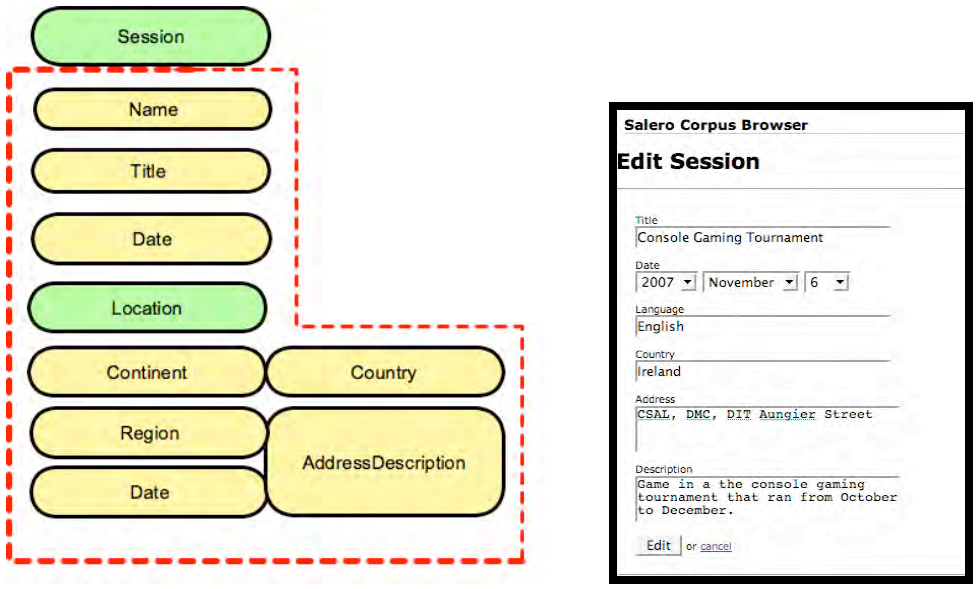


Figure 22: The session sub-schema as defined in the IMDI schema containing various details about the session and (a), and how it was implemented in the backend database (b).

The session sub-schema was implemented with the session title, date, language, country, address and a description of the session. The continent and region elements were omitted, as were the session name and location elements; the title element was sufficient for naming the session and the Description field provided for sufficient location information. A language element was also added to specify the language of the session adding a greater degree of flexibility: it was possible to conduct a session

in a language different to that of the participants. This is advantageous for other research, e.g. studying the emotional parameters of non-native English speakers engaged in a task where English is the communicative language.

6.2.3 Participant

Within a session, the definition of a participant is a useful aspect of the IMDI standard, as it allows the various participants in a speech recording to be documented for later consideration. In many instances, participant details need to be vague and non-specific to ensure that ethical standards are adhered to by preserving anonymity and this can be given as an option for each testing participant. One can choose to implement all aspects of the participant element or leave some out if they are not necessary. However, as mentioned, more detailed participant information would be of use for certain types of queries: while it would be advantageous to collect and store all participant data, there are times when information regarding participants is required and/or requested to be kept anonymous. The implementation of unused elements from the outset or at a later date is dependent on the type of corpus being created and is at the discretion of the corpus architect. This is an important consideration as some participants might only consent to their participation if they remain anonymous, negating the need for all elements to be implemented; the flexibility of the IMDI schema allows extra elements to be retrospectively added as the corpus grows or as they become necessary. In addition, the keys element allows for user-defined elements to be added as part of the participant sub-schema in the same manner. Figure 23 illustrates the participant sub-schema and its instantiation in the backend database.

(a) (b)

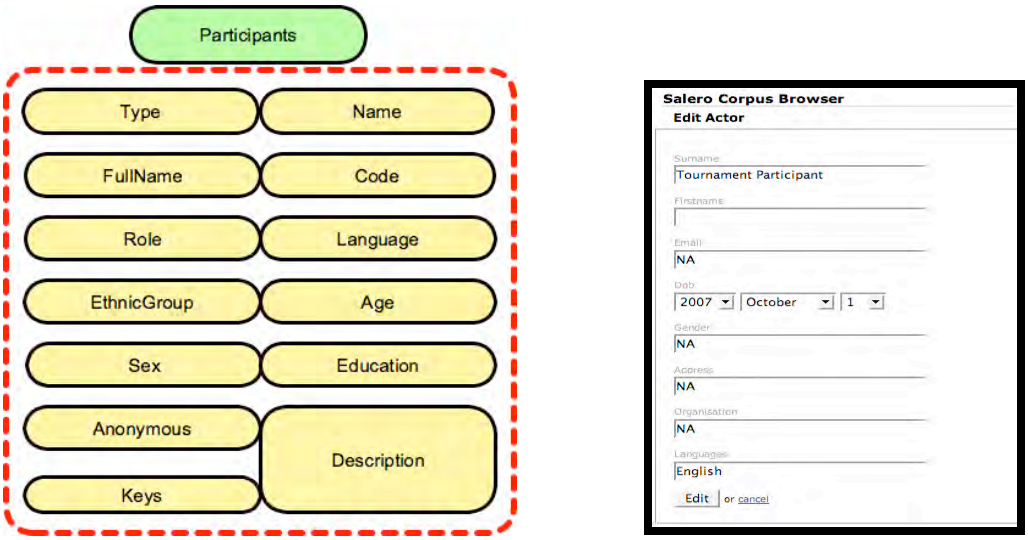


Figure 23: How the participant element is defined within the IMDI schema, containing detailed information about the participant (a), and how it was modified and implemented in to the backend database (b).

The participant sub-schema was implemented with some elements omitted and others added. The Name and FullName elements became surname and first name, age became date of birth (this was more precise than just age), and gender and language were preserved. Entering a code for the name in the Surname and First Name fields preserved the anonymity of participants. Each participant was given the surname 'participant' with the first name being a number and the first letter of their first name e.g. John Smith would become '12(J) Participant'. The code given to each participant is noted in the consent forms they signed and kept in a separate excel spreadsheet. Only the collector and creator of the project knew the full identities of the participants.

6.2.4 Content

The content metadata relates to specific activities/recording sessions within a session. The metadata defines the exact nature of the activity, defining the genre and sub-genre (task), the modality, the language, and a description of the content. User defined elements can also be added; thus tailoring the content sub-schema to MIP based content. The genre and sub-genre definitions can be implemented as open vocabularies, while the other terms remain standard IMDI closed vocabularies. Each session has content of some description and content metadata can be applied to more

than one session within a project. This allows a number of different recording sessions to take place within a project using the same type of content. The content element is of particular use in annotating assets obtained using MIPs, especially a task-based, success/failure social interaction MIP (see 4.2.4 and chapter 7 and 8); the open vocabularies and the user defined elements (keys) allows for the implementation of MIP relevant descriptors within the content element. This is one of the main advantageous in using the IMDI metadata schema. Figure 24 illustrates the content sub-schema and its instantiation in the backend databse.

(a) (b)

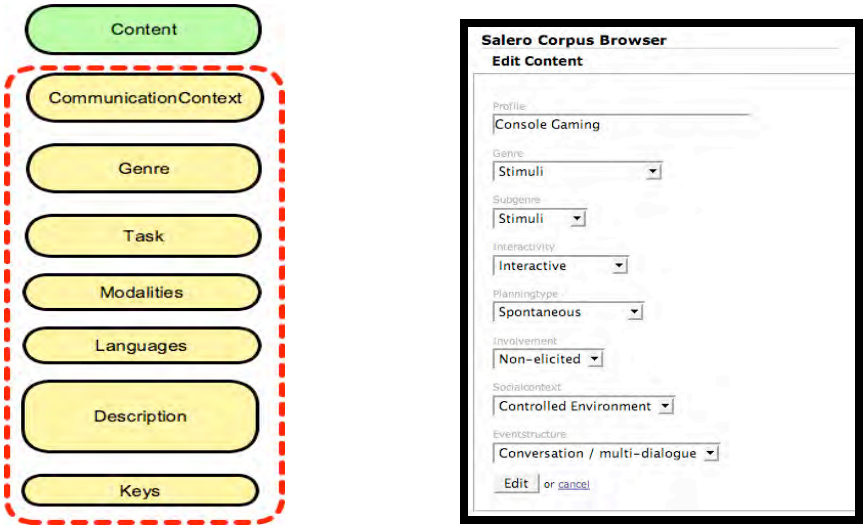


Figure 24: How the content element is defined in the IMDI schema (a), and how it was implemented in the backend database (b).

The content sub-schema was implemented with a number of extra elements included, specifically in relation to MIPs. CommunicationContext was implemented as Profile to name the relevant content. The genre element was implemented with 11 different options available and the task element became the sub-genre option. Since the corpus contained audio only, the modality element was omitted, as was the language element (being a part of the session sub-schema instead). The description element was also omitted and a number of extra user defined elements were included: the level of interactivity, the level of spontaneity, the social context and the type of communication that took place. These extra elements were necessary to more accurately describe the MIPs used, as the existing IMDI content sub-schema was inadequate in this regard.

6.2.5 Incorporating The Adapted Schema In To The Backend Database

The adapted project, session, content and participant sub-schemas were all incorporated into the database. The IMDI sub-schemas were instantiated as separate screens of the backend database interface. Figure 25 illustrates how the four metadata sub-schemas are related to the main asset interface screen within the database. The instantiation of the metadata schemas was structured to allow metadata to be entered prior to, or during the upload of an asset. Metadata previously entered became available at the upload stage via the Ajax autosuggest facility. If the relevant metadata had not been entered previously, then it was able to be entered during the upload process and would then be available for reuse as with the previously entered metadata. Incorporating the addition of metadata into the upload process reduced the number of steps needed to enter data, allowing for a certain degree of flexibility in the data entry procedure. The ability to upload a single asset or a batch of assets was implemented allowing metadata to be applied to groups of assets with the same metadata (e.g. a group of assets from the same session, by the same participant, with the same content), with the Ajax autosuggest facility being used in both cases.

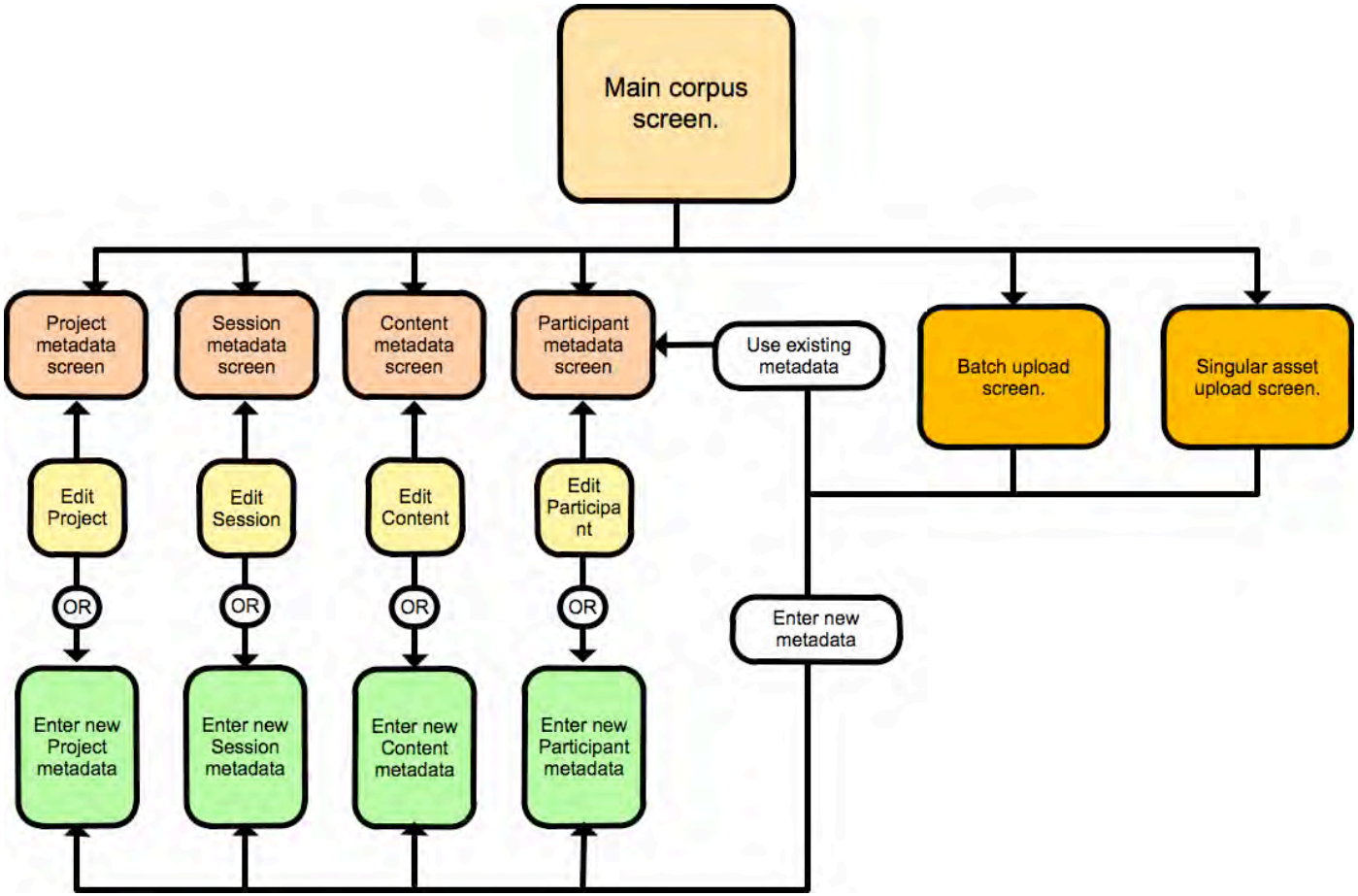


Figure 25: Illustration of how the adapted IMDI sub-schemas are incorporated into the backend database. New metadata can be entered separate to or as part of the asset upload procedure. Existing metadata can be edited once entered. Existing metadata cannot be edited during the upload procedure

6.3 Acoustic Analysis And Annotation

Assets that are to be used in the corpus need to be processed and annotated prior to their upload. The recordings from the MIPs (chapters 7 and 8) were segmented into smaller speech assets, as it was important to ensure that listeners were more focused on *how* something was said as opposed to *what* was said. It is possible that some listeners might evaluate the emotional content of the speech differently once the context within which the speech took place was known (Iida 2003)**.** The emotional trajectory of a speaker may change over time; a person may start a conversation in one emotional state and transition into other states during the conversation (Cowie, Douglas-Cowie et al. 2000): the use of short speech assets are an attempt to focus on singular states that are part of a speakers overall emotional trajectory and thus help avoid instances of overlapping emotional states. Analysing assets is simplified when dealing with singular emotional states and research has suggested that humans can recognise emotions in short speech utterances of at least 1.5 seconds in length (Shami 2006; Mansoorizadeh 2007).

The audio recordings collected from the MIPs were converted to the Redbook wav standard (IEC 1999) (4.4.2) and manually segmented into assets using Pro-Tools (Avid-Technology 2007). Most assets were 3-10 seconds in duration and were segmented to capture single, complete utterances by participants. The assets were processed using the LinguaTag application to test the parsing of the XML SMIL file into the backend database (Cullen, Vaughan et al. 2008a) (5.2.2)**.** The wav versions of the files were stored in the backend database for archiving purposes. Mp3 versions of the wav files were created for use in the listening tests (5.3 and 6.4.2). The original, un-segmented, HD quality (192 Khz/24 bit) audio recordings were stored on the database and on a separate server for archiving and backup purposes.

The LinguaTag SMIL XML file for each asset was uploaded along with the corresponding wav and mp3 audio files. The information within the XML file was parsed by the database into tables. The XML file, the wav and the mp3 files were stored in local folders and maintained a connection to the asset information in the database. The LinguaTag XML file was successfully parsed in all cases and

demonstrated an ideal method for annotating the acoustic parameters of speech within a speech corpus. Each asset had a connection to the three corresponding files and the parsed information. Figure 26 illustrates how each uploaded asset maintained a connection to the three uploaded files and the parsed information. These connections simplified the analysis of the assets, both emotionally and acoustically. The emotional rating of assets using a dimensional rating tool can only take place once the assets have been uploaded, while the parsed LinguaTag data can be queried using standard SQL commands. The LinguaTag analysis and resulting data is only one aspect of the acoustic analysis process, other parameters need to be considered (3.2) and can be measured using PRAAT (5.2.1). However the PRAAT data cannot be parsed into the database in the same way that the LinguaTag data can and so this aspect of the analysis must be completed manually. Future work will consider a method of combining the acoustic analysis abilities of PRAAT with LinguaTags ability to produce an audio analysis file that can be parsed into a backend database (10.3.4).

Figure 26: Illustration of how each asset links to the three uploaded files stored on the database and the parsed LinguaTag data. Each piece of data is linked to the asset within the database. Each entry in the database has a corresponding mp3, wav and XML file.

6.4 Emotional Dimensional Annotation

Having examined the advantages of using a large listening group to obtain emotional dimensional ratings in 5.3, consideration must now be given to methods of collecting these ratings from a large listening group and storing them alongside the metadata and acoustic data. While most listening tests take place in a laboratory setting over a long period of time (usually a few hours for each listener is required), this method is impractical for large listener groups. While the laboratory setting allows for a high degree of control to be exerted over the listening tests, it is possible that it colours the judgement of the listeners by giving rise to demand effects. A balance needs to be struck between the level of control that a laboratory setting allows and the increased listening group size that a non-laboratory setting allows for.

6.4.1 Web-Based Technologies As A Means Of Carrying Out Large-Scale Listening Tests

There are various technologies that can be used to carry out online listening tests. Web browsers are an ideal delivery platform and interface for a listening test, as they are the most widely used method of interacting with the World Wide Web; every computer with web access will have a web browser of some description. There are numerous applications and software platforms that can be used to develop web based applications and interfaces. Adobe Flex is an open source framework, used to create expressive web applications that are compatible with all major web browsers and operating systems (Adobe 2009). It can be used to create web or desktop based applications and is part of the wider Adobe Flash Platform. The Flash player browser plug-in is needed to view or use Flash interfaces, and is installed by default with all major browsers. It is one of the most ubiquitous pieces of software available, thus ensuring that the majority of web users are able to use Flex applications. The Flex programming platform works over two main screens: a design screen were graphical and user input elements can be laid out on canvas, and a code screen where Action Script code can be written to add functionality to the graphical elements (Noble 2008).

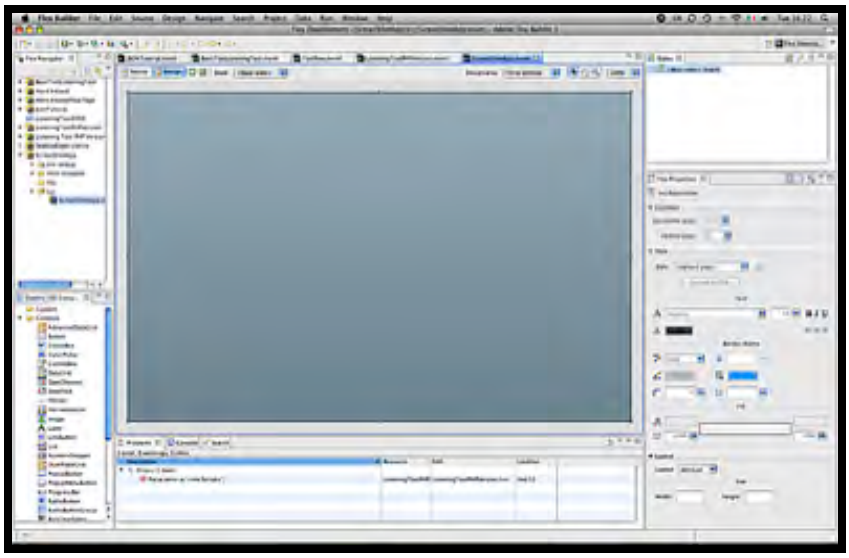


Figure 27: The Flex 3 design interface showing the main canvas screen were application elements are laid out.

Flex applications can be web based or desktop based, with Flex also enabling the creation of applications for various social networking sites such as Facebook. With such a large amount of people using social networks like Facebook[24] (and Twitter), Flex provides a powerful means by which to develop a tool to carry out large-scale listening tests.

6.4.2 Developing An Online Listening Tool For Large Scale Listening Tests

The emotional annotation can only take place once assets have been uploaded to the database; this is the only viable way to access an organised set of assets and store their subsequent ratings. Considering that most listening tests use only a small amount of often-expert raters, online methods offer distinct advantages (5.3.1). Harnessing the power of the now ubiquitous World Wide Web allows for a large number of ratings to be achieved in a relatively short space of time. To this end, a web based online listening tool was created utilising Adobes Flex platform. Due to the highly structured nature of the back-end database, the tool was able to connect to the database, extract

[24] The latest figures listed on the Facebook site claim that Facebook has more than 350 Million users, with each user having an average of 130 linked friends. Keeping in mind the pace of technological change, especially in relation to social networking sites, these figures are likely to change on a regular basis (Facebook 2010).

ten random assets for rating and write the ratings back to the database for future analysis.

6.4.3 Design And Implementation

The tool is designed to have a clean interface and be easy for users to understand and use; all aspects of the tool existed on a single interface screen. There were a number of considerations that were implemented within the tool:

1. Users were able to replay the audio as many times as they wanted.
2. A visual indicator that audio was playing was included: if there were problems with the audio settings on the user's computer this would indicate that audio was playing.
3. The number of assets presented to a listener was limited to a small set with the option to rate more once these were rated. This helped focus a listener's attention: presenting too many assets could have led to boredom and potentially spurious ratings.
4. An option for skipping an asset if a user did not want to or could not rate it was included.
5. A two-dimensional model of activation and evaluation, similar to the FeelTrace tool (2.5) was implemented as two separate sliders.

Figure 28: The main listening tool interface screen showing the two dimensions of the circumplex model as separate sliders on the right and the audio controls on the left.

As discussed in section 6.4.1, the Adobe Flex platform is ideal for creating Rich Internet Applications (RIAs) and provides an excellent platform for creating a browser based listening tool. A simple interface was created (Figure 28) with Action

Script 3 code being implemented to add functionality to the buttons and to retrieve asset mp3 files from the database (Appendix B).

The interface is split visually into two sections, left and right. The left hand section contains a play button, a simple graphic that responds to the playing audio and a row of ten numbered squares.

Figure 29: The visual graphic indicating audio is playing and the numbered line to indicate how many of the 10 random assets have been rated

The right hand side consists of two horizontal sliders, an active/passive (activation) slider and a positive/negative (evaluation) slider. Both ran from 0 (on the left) to 10 (on the right) in 0.5 increments, giving a total of 21 different points on each dimension:

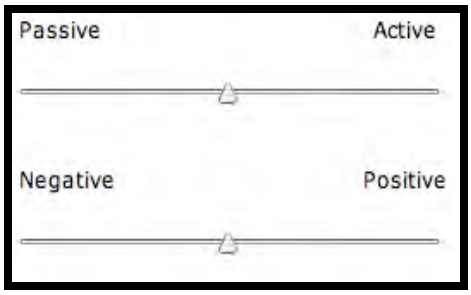


Figure 30: The two ratings sliders were the separated dimensions of the circumplex models. The activation dimension is at the top and the evaluation dimension is at the bottom.

These sliders are the dimensions of the circumplex model separated out into two individual sliders. The circumplex model was not implemented in its original state as it was felt that presenting the full circumplex model to users would have complicated the interface and would have necessitated a longer explanation of how to use the tool. Similarly, with their implementation of the circumplex model, Sánchez and

Hernández et al. found that presenting the complete model to users overburdened them with information; their solution was to use a simplified version of the model that presented the user with the information in a number of steps (Sánchez, Hernández et al. 2006). As a result, the separate slider implementation was executed to simplify the interface and make it quick and easy for users to understand; it also ensured users used the two sliders independently of each other, thus giving due consideration to each dimension. This also facilitated the analysis stage of the research as the acoustic correlates of each asset could be analysed in relation to the ratings received on each individual axis (chapter 9).

The Rate button is clearly visible, with the greyed out Do Not Rate Button only becoming active once one of the sliders is moved: this ensured that users could not keep clicking the 'Do Not Rate' button to skip assets unless they had made some effort to rate them. A simple four-screen introduction to explain the purpose of the test and instruct users on how to use the tool was also created (Figure 31). Small-scale user testing, in the form of conversational feedback, was carried out to investigate whether users understood the instructions and how to use the tool correctly. The testing group of ten people responded positively, with some suggestions made regarding the instructions on the second introduction screen.

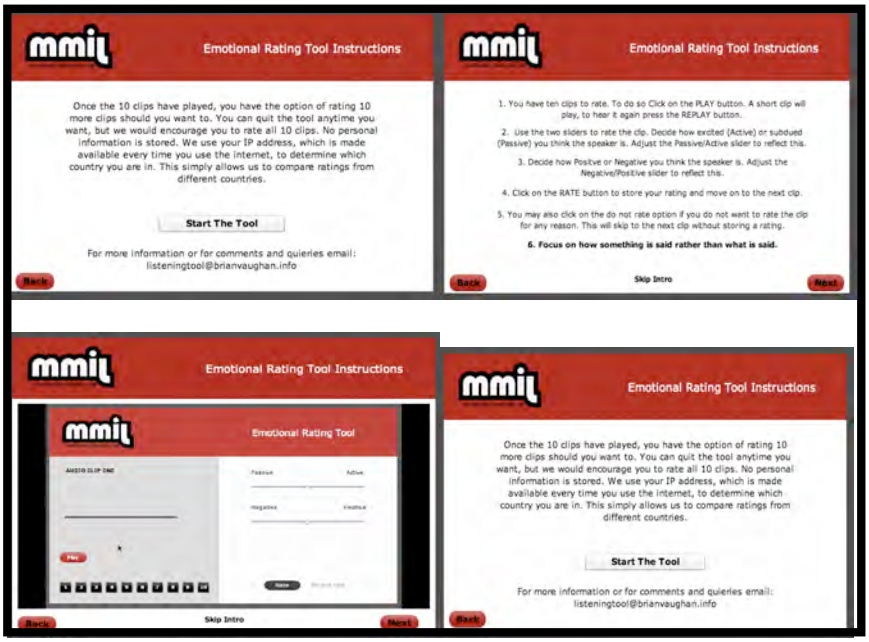


Figure 31: The four-screen introduction to the listening tool. The screens explain how to use the listening tool and its purpose in a clear and straightforward manner.

It was necessary to randomise the assets that were presented to listeners in the listening tests used to rate the assets from the MIPs in chapter 7 and 8. It is possible that the order in which assets are presented to listeners will have an effect on the ratings received. There is no way of determining this aside from comparing listening tests using ordered sets of assets to a randomised listening test. Any other approach would constitute an *a priori* categorisation of the assets thus defeating the purpose of carrying out listening tests.

When the tool is first accessed, 10 random assets are loaded into the tool. As the user rates each asset, a square on the number line becomes greyed out to show their progression. Each time an asset is played, the waveform graphic responds to indicate that audio is playing.

6.4.4 Corpus Visualisation

In order to provide a visual interface that could represent the data in the corpus in an easily understood manner, a prototype interactive, real-time, corpus visualisation interface application (CorpVis) was built (Cullen 2009). The interface displays basic information about the assets in the corpus: gender, number of assets, pitch and intensity information and displays all the assets on a circumplex model according to their emotional rating. This allows a user to have an overview of the corpus and the assets within it along with a visual representation of the emotional ratings. The visualisation tool queries the various pieces of annotation data within the corpus. Whilst only a prototype, it demonstrates how the annotation data can be utilised in a meaningful way: providing an intuitive and interactive method of viewing the complex data within the corpus. The interface consists of two screens: the main overview screen; giving an overview of some of the demographic, acoustic and emotional information in the corpus; and a more detailed acoustic information screen giving more detail regarding some of the acoustic parameter values for a selected asset.

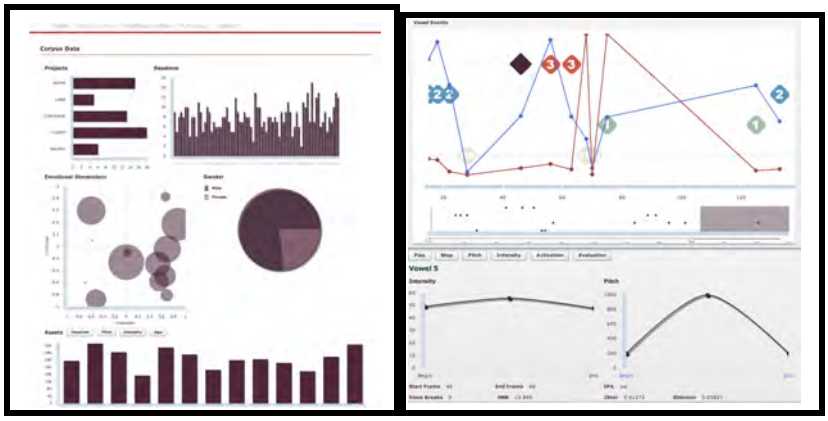


Figure 32: The two main CorpVis screens, (a) is the main overview screen showing basic demographic, acoustic and emotional data and (b) is the more detailed acoustic information screen showing all the acoustic data related to each asset.

CorpVis was integrated into the wider corpus structure (Figure 34), alongside the emotional rating tool (6.4). Future work will examine methods of improving the visualisation interface and increasing the level of user interactivity (10.3.4).

6.5 Asset Upload Procedure

Once the adapted IMDI sub-schemas and acoustic annotation methodology had been instantiated in the backend database, and the listening tool deployed within the wider corpus framework (see Figure 34 for an overview), assets were then uploaded, either as a batch, or as a single asset upload. The batch and singular upload procedure was as follows:

1. If assets were to be uploaded in a batch, the corresponding wavs, mp3 and SMIL XML files were placed in a folder. These assets were to have the same IMDI annotation data applied (e.g. a number of assets from the same actor, in one session, of the same content).
2. The folder was compressed into a ZIP file with the same name as the folder.
3. If uploading a singular asset, the asset and its XML file were selected using two separate file upload dialogue boxes.
4. Using the upload interface the file(s) (ZIP or asset and XML file) were selected and the annotation data applied via a number of drop-down boxes.

New annotation data could have been created at this stage if project, session, actor and content data had not been created previously.

5. The files were uploaded and stored in the database and the data contained in the SMIL XML file were parsed and written into tables in the database along with the IMDI annotation data. This data was displayed via a HTML administrator interface and the public CorpVis interface. Figure 33 shows the two separate upload screens.

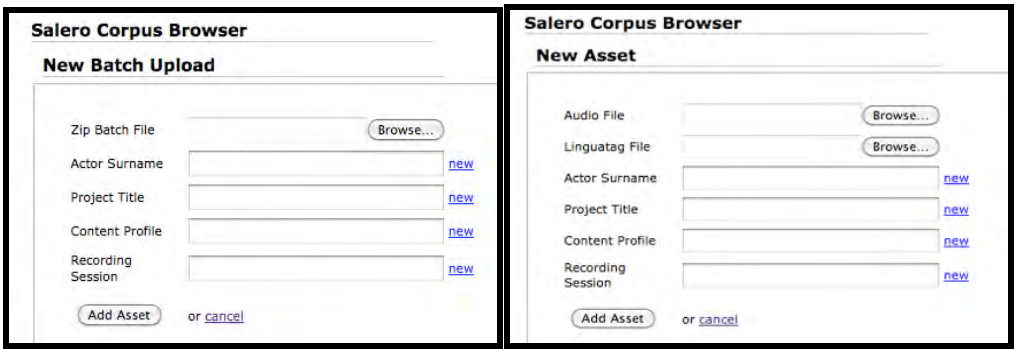


Figure 33: The batch upload and singular asset upload screens. These upload screens are incorporated into the wider database structure.

Figure 34 illustrates the structure and connectivity of the database, and demonstrates the practical instantiation of the three-tiered annotation approach, conceptually illustrated in Figure 16, in chapter 5.

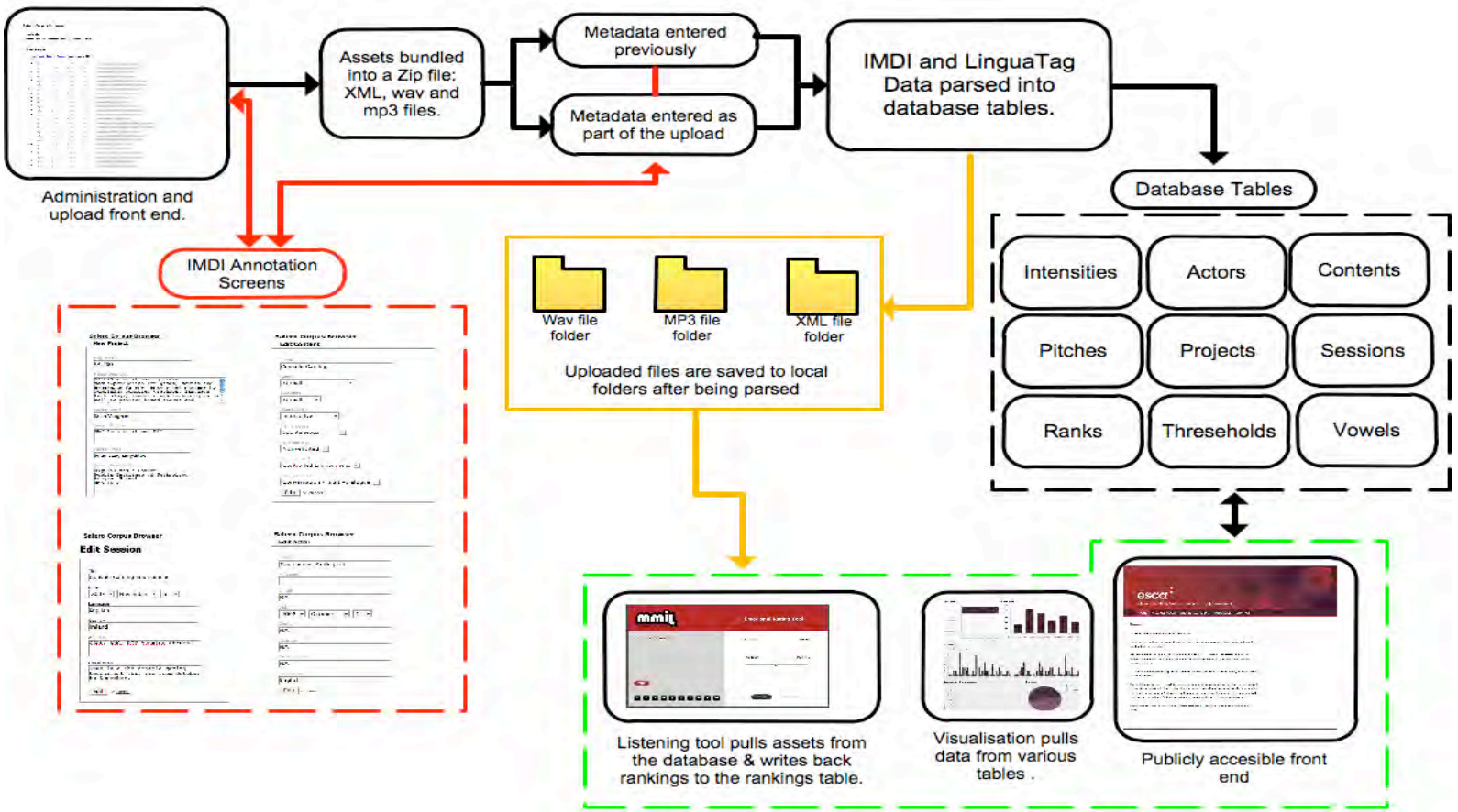


Figure 34: Overview of the structure and connectivity of the backend database illustrating how the various levels of annotation are integrated into the database structure and upload process. The green outlined section is the public front end through which users interact with the corpus assets. The red outlined section is the IMDI metadata creation screens.

6.6 Discussion

Taking the annotation approach discussed in chapter 5, and instantiating and structuring it logically in a backend database ensures that the usability and usefulness of the speech corpora is extended beyond its original remit. It also enables the data contained within the corpus to be accessed using external applications and research tools. The development of powerful emotional speech applications and the in-depth analysis of emotional speech require ever larger amounts of data, to overcome problems such as data-sparsity and to enable use of the most appropriate data .

Current database and web technologies are powerful tools to be utilised in the creation of speech corpora. Web based database technologies like MySQL allow corpus assets to be structured coherently while being easily accessible by third party applications and specially developed Rich Internet Applications (RIAs). PHP and Ruby-on-Rails are two popular web technologies that enable data stored in a backend database to be utilised to its full extent, with Ruby-on-Rails allowing for rapid development and deployment (6.1). A persistent backend database is conceivably the best method of structuring and actualising a three-tiered annotation of ESAs while addressing issues of data-sparsity and reusability.

While various metadata schemas exist, the IMIDI schema is arguably the most suited to the creation of an emotional speech corpus, especially when MIPs are used to obtain assets (5.1.1 and 6.2). The sub-schemas within it, and their related elements, are ideal for describing MIP based assets, particularly a task-based, success/failure, social interaction MIP. The flexibility of its elements allows it to be adapted to the specific needs of a corpus while retaining its overall structure. Incorporating the adapted IMDI schema into the upload process introduces a level of flexibility in the data entry aspect of the corpus and allows the MIP used to obtain assets to be concisely detailed (6.2). While the IMDI schema was the most suited out of the metadata schemas examined, a certain level of adaptation was necessary to suit the needs of the corpus, even though the schema had some user-defined elements built into it.

While PRAAT and LinguaTag can both be used to carry out acoustic analysis, PRAAT is capable of a more extensive acoustic analysis than LinguaTag. However, as previously noted (6.3), only the LinguaTag data can be parsed into the backend database and PRAAT must be used to carry out manual or semi-automatic acoustic analysis. While it may seem that a complete acoustic annotation methodology and implementation is necessary, the converse may actually be advantageous. Implementing an acoustic annotation methodology would more than likely require the determination of the set of particular acoustic parameters related to speech analysis (emotional or otherwise). In chapter 3 it was demonstrated that there is a consensus on certain acoustic parameters in relation to certain emotional states, but this does not preclude the examination of other parameters that have not received as much attention in the literature. It may well be more advantageous to instantiate an acoustic annotation schema that incorporates a certain level of manual or user defined analysis and annotation. An acoustic annotation schema that allows for the instantiation of user-defined annotation elements in a similar manner to the adapted IMDI schema would allow for flexibility in an acoustic analysis and annotation methodology. Future work will consider the development of a flexible acoustic annotation schema (10.3.4).

The final tier of the proposed three-tiered approach is the emotional annotation of the ESAs. As discussed (2.5), the circumplex dimensional model allows the emotional content of speech to be rated on a two-scale axis. This avoids the use of ambiguous and subjective emotional terms and categories: it can be argued that there is no consensus in the field of emotional research regarding a definitive list of emotion terms and categories. A dimensional model provides a simple and straightforward method of rating emotional content (2.5). Listening tests have previously been used to determine the emotional content of speech and mainly use a list of emotional terms/categories (5.3). However, the majority of the listening groups used are small with an average of 28 people per group. Only in a few cases have larger listening groups been used. Large listening groups are advantageous as they, reduce the likelihood of finding chance relationships between variables, ensure that the probability of random deviations from the sample mean decreases and avoid ‘best fit’ errors. This allows a more confident set of statistical results to be obtained (5.3.1). Listening tests normally take place in a laboratory setting, a limiting factor in the size

of the listening group and one that could possibly lead to demand effects; online listening tests offer the advantage of reaching a potentially huge listening group without users feeling constrained by a laboratory setting (6.4). Adobes Flex programming environment provided a means of creating a user friendly, browser based listening tool. Flex can also be used to create plug-ins for social networking sites, allowing users to access a listening test via an oft-used online resource (6.4.1). Harnessing the power of the interconnected social aspects of the web offers the potential to reach large numbers of people for listening tests. Social networks such as Facebook and Twitter have millions of interconnected users, and allow for the mass dissemination of information in a potentially rapid manner. Figure 34 illustrates the connectivity and practical realisation of the three-tiered annotation methodology within the backend database.

6.7 Conclusion

This chapter considered the annotation of emotional speech assets (ESAs) based upon the three-tiered approach discussed in chapter 5. The technical aspect of bringing together the three forms of annotation in a coherent structure was discussed, contending that a robust backend database was needed implement and structure the three-tiered annotation methodology (6.1). Various technologies exist to enable this: MySQL, PHP, Ruby-on-Rails and Ajax were discussed, with Ruby-on-Rails offering the advantage of allowing complex web based applications to be created quickly and efficiently, especially in conjunction with MySQL (6.1). A backend database, structured in a logical and coherent manner allows the various annotation data to be searched and queried as needed, thus providing a powerful and interactive speech corpus. The implementation of adapted IMDI sub-schemas into the database structure was discussed, examining the incorporation of the metadata data entry into the asset upload process (6.2.5). This allows metadata to be entered prior to or during the asset upload process, thereby introducing an element of flexibility into the metadata entry procedure (6.5). The implementation of the acoustic annotation was next considered and discussed the segmentation, conversion and storage of the original high quality audio assets; it was argued that the original recordings needed to be segmented to remove the context within which the speech took place and to focus on singular emotional states (6.3). The point was made that only the limited acoustic analysis of

LinguaTag was able to be parsed into the backend database and that future consideration should be given to a more comprehensive and flexible acoustic annotation schema that incorporates user defined annotation elements (6.3). The annotation of emotional dimensions was next considered and detailed the design and implementation of a browser based listening tool to carry out large scale listening tests (6.4.1 and 6.4.2). This was followed by an examination of an interactive corpus visualisation interface that presents a user with an interactive, aesthetically pleasing and informative visualisation of the emotional dimensions of the corpus assets along with details of some acoustic parameters (6.4.4). The visualisation interface was an example of how a corpus underpinned by a persistent backend database can be connected and utilised via web-based tools.

This chapter contributed to the answering of research questions: RQ 2, RQ 3, RQ 5 and RQ 6.

7. Case Studies: Developing An MIP To Elicit Natural Emotional Speech

In order to examine the more subtle underlying emotional states (2.3, 2.4 and 3.6.1) a method of obtaining emotional speech needed to be devised: this chapter details the development of a task-based false feedback social interaction MIP to elicit natural underlying emotional speech. Two case studies were carried out to ascertain the most effective method for obtaining natural emotional speech in a high quality audio environment; a balance was needed between audio quality and the induction of natural underlying emotional states. Both case studies were a combination of the success/failure MIP and the social interaction MIP (both from MIP group 4) incorporating aspects of the gift giving MIP (MIP group 3), and used isolation booths and high quality audio equipment to record the participants' speech at a professional HD audio standard (4.2.4, 4.3). Induction of emotional states was achieved through placing participants in a socially cooperative or oppositional situation, hindering or aiding the participants in the attainment of a set goal and giving false feedback regarding their performance. The first case study used Kehreins original Lego experiment (Kehrein 2002) and the popular Tetris game to test the use of sound booths and HD recording equipment, as well as methods of externally manipulating the experimental conditions. A second case study using contemporary console games built upon the first case study and investigated the use of computer games as a method of mood induction (4.2.8). The experimental setup in both case studies resulted in the recording of clear, noise free high quality naturalistic underlying emotional speech assets. The recordings from the gaming MIP were segmented and uploaded for rating using the listening tool (6.4 and 6.5). The number of ratings received was not enough for any meaningful statistical analysis to be carried out. This led to a refinement of the methods used to obtain ratings for the final shipwreck MIP detailed in chapter 8.

In each case study, the basic setup and equipment used is first discussed, detailing the hardware and configuration. The experimental design is then discussed explaining the different experimental conditions before the pros and cons of each case study are examined.

This chapter contributes to the answering of research questions: RQ 1, RQ 2, RQ 3, RQ 4, RQ 5 and RQ 6.

7.1 Ethical Considerations

Various ethical considerations must be taken into account in eliciting emotional speech from participants. Inducing emotional states in participants must be done with care; inducing full-blown/primary emotions from participants may be ethically dubious (2.3, 2.4 and 4.2.9) (Campbell 2000). These are powerful and complex states that are not considered to be constituent elements of the majority of human interaction, often interrupting the communicative process (2.3). Secondary or underlying emotions are considered constituent parts of human interaction and are not as intense and disrupting as primary or full-blown emotions and thus their induction is not as ethically dubious as that of primary emotional states. Their central role in the human communicative process suggests that their examination may of greater relevance (Cowie, Douglas-Cowie et al. 2000). All participants signed a consent form prior to any experiment (Appendix C): the form and the experimental conditions of all experiments were reviewed and passed by the DIT's ethics committee. The form clearly stated that participants would have their speech recorded for research purposes but the exact nature of the experiments was not mentioned; the research was described as speech research and considering the argued importance of emotion in human communication, this was the partial, but not the whole, truth. It was vital to the success of the experiment that participants were not aware of the intent to elicit emotional speech and the misleading of participants was necessary to avoid demand effects (4.2.7)

7.2 First Case Study: Lego and Tetris

Researchers have had success using isolation booths in audio-based experiments (Kehrein 2002; Ramus 2002; Kooijman. V 2005). Isolation booths have three advantages in speech experiments: the participants are not distracted by external noise or activity, the recorded audio is free from external and unwanted noise and they ensure the dialogue between participants can be recorded as two separate audio streams. Recording the audio as two audio streams makes subsequent analysis easier and avoids cross-over as well as overlapping speech elements (Cullen 2006). For

every MIP experiment, each booth was equipped with a flat screen monitor, a pair of headphones, a microphone stand and a professional Neumann microphone (Neumann 2009). For this case study, the microphone was connected to an Apple Mac-based (Apple-Computers 2009) Digidesign Pro-Tools Mbox2 recording system (Avid-Technology 2007) (4.4). The audio signal was digitised at 96kHz/24Bit (4.3) and recorded using Pro-Tools software as two separate audio streams. A second machine was connected to the two monitors in each booth to run the Tetris software. For the Tetris experiment a keyboard was placed in one of the booths and connected to the external machine running the software. Light in the booths was provided by a string of white LEDs arranged around the internal edges of the booth.

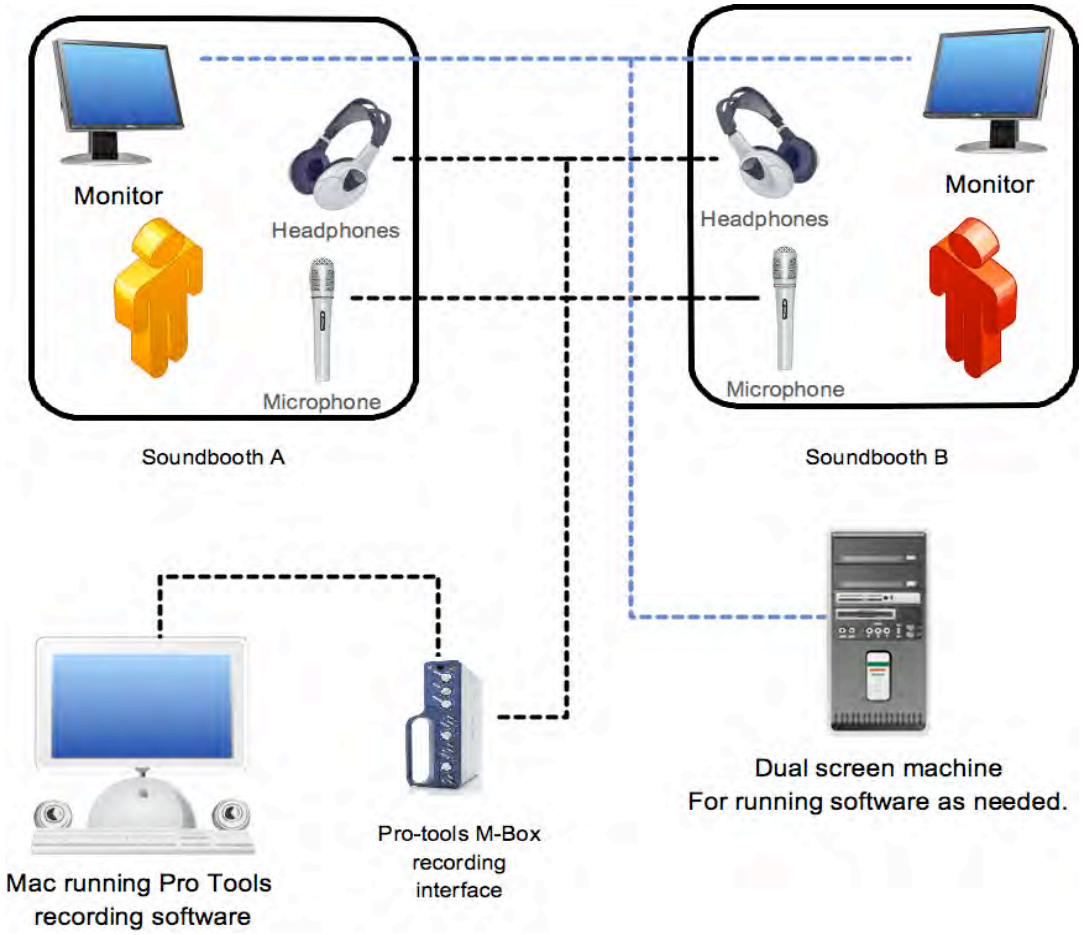


Figure 35: Basic equipment setup for the Tetris experiment. Participants sat in soundproof booths with a monitor, microphone and headphones. LEDs around the inner perimeter of each booth provided light. A Pro-Tools Mbox 2 recording system was used for the first case study recordings.

7.2.1 Experimental Design For The Lego And Tetris MIPS

Using the equipment setup, two experiments were designed in order to elicit emotional responses from participants. The experiments used Kehreins Lego construction task (Kehrein 2002) and the popular Tetris computer game. Both the Tetris (7.2.3) and the Lego (7.2.2) experiments consisted of two participants engaging in a cooperative-based task while their dialogue was monitored and recorded externally. The Lego game mainly used missing pieces to hinder the participants as a method of manipulation. The Tetris based game was designed so that various elements of the experiment could be externally manipulated in order to engender emotional states within the participants. A small reward was offered in each experiment for the successful completion of each cooperative task, providing an added impetus for completing the task within the set boundaries. This reward was offered for successful completion of the set tasks, for either beating a certain high-score or completing a task in a certain amount of time. Manipulation of the experiment ensured that these goals were rarely attained. However the use of a small cash reward has been found to be a successful method in engendering positive emotional responses (elation) in participants (4.2.3). Participants taking part in the experiments were not allowed to bring any devices that displayed time into the booths, thus allowing false information to be given regarding how much time was left or had elapsed.

7.2.2 Experiment 1: Lego

This experimental design was first used by Kehrein in 2002 (Kehrein 2002). Two participants in isolation booths had to cooperate to build a Lego construction. One participant (participant A) gave instructions on how to build it and the other participant (participant B) followed these instructions. The participant building the construction was the only one who had access to the pieces of Lego needed to complete the structure and had to follow the instructions as closely as possible to complete the task. Necessary pieces of Lego were removed beforehand to hinder the participants.

For this experiment a Lego fire engine was used (Figure 36). The main aim of this experiment was to test the overall experimental setup and was a preliminary experiment to test the manipulation of the experimental conditions. It was found that,

while removing pieces did hinder the participants thus eliciting emotional speech, a more direct and controllable form of manipulation was needed. Participants adapted quickly to the fact that a piece or pieces were missing. Removing too many pieces also led to participants realising that there had been deliberate manipulation. These observations led to the development of a more controllable Tetris experiment that did not rely on manipulating the conditions prior to the experiment commencing and provided an improved means of manipulation.

Figure 36: The Lego fire engine used in the Lego case study experiment.

7.2.3 Experiment 2: Tetris

This experiment used the hugely popular Tetris puzzle game[25]. Tetris consists of seven different shaped blocks that have to be stacked and slotted together, like a vertical jigsaw, in order to complete a continuous horizontal row of blocks with up to four complete rows being possible in any one instance. Players score points for each complete row. Each of the seven blocks can be rotated clockwise and counter-clockwise and moved left to right across the screen. Once a player is happy with the position of a block it can be moved straight down into position, otherwise it moves slowly downward until it comes to rest on another block or the bottom of the screen.

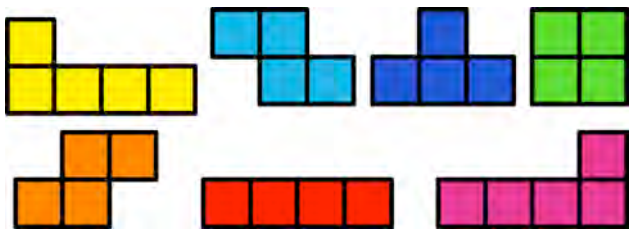


Figure 37: The seven Tetris shapes used in the Tetris game. Each block can be rotated an fitted against the other blocks in a variety of ways.

[25] http://www.tetris.com/ Home page of Tetris. Details about its history and popularity

For the purposes of this experiment, participant B manipulated the blocks according to the instructions given by participant A. Participant A instructed participant B on how to rotate, move and stack the blocks with both working together in order to achieve an agreed score within a certain time frame. Participant B was not able to see how the blocks were stacking up but could control their movement, while participant A was able to see the game and the actions carried out by participant B but had no control over the movement of the blocks. Thus cooperation was essential for the completion of the task. Figure 38 illustrates how the experiment worked.

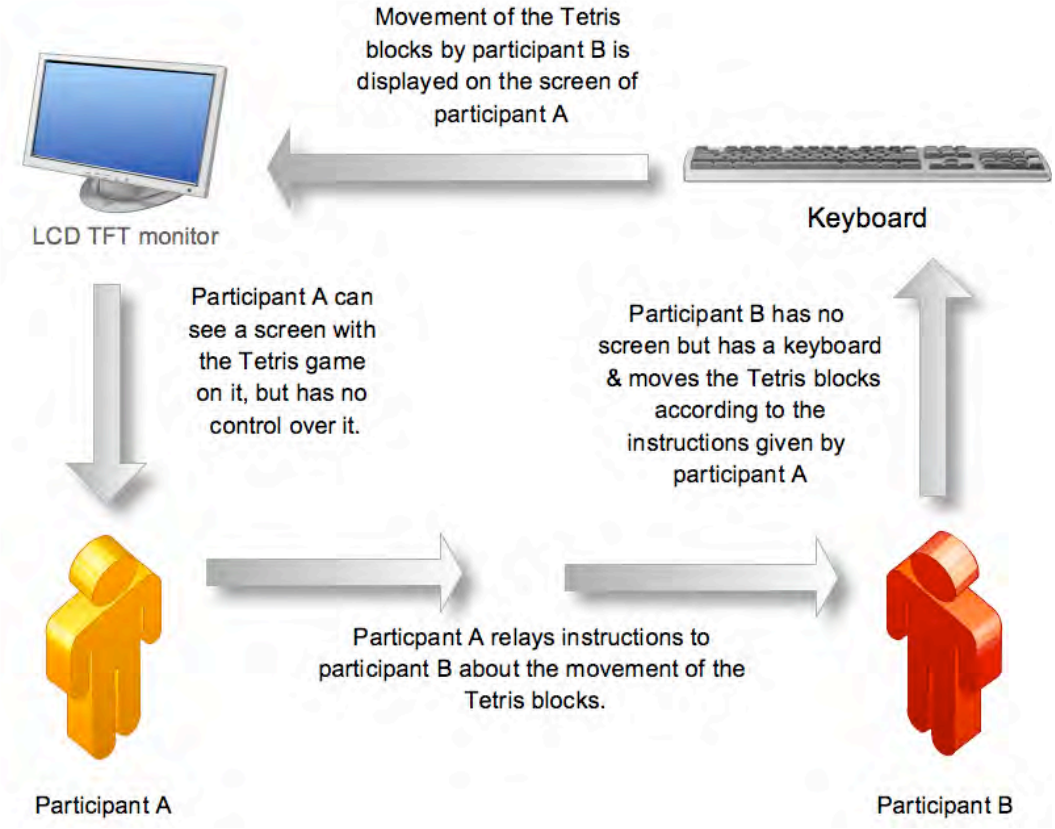


Figure 38: Diagram showing how the Tetris experiment was setup. Participant A can see the Tetris game on a screen and gives instructions to Participant B who can not see the game but can control the movement of the Tetris blocks according to the instructions given by Participant A.

The co-operative process was hindered by preventing participant B from hearing the instructions properly (through cutting the audio feed to the booth), by altering the time allowed to achieve the desired goal, and by externally controlling the falling pieces without either participant knowing (via an external keyboard). The participants were initially aided by giving a generous amount of time to complete the task and by lowering the score that needed to be attained. The Tetris game used was a basic Flash

based version, freely available in numerous locations on the Internet. A cash reward (4.1.3) was offered for the obtainment of a certain score, which was specified by the researcher.

7.2.4 Results Of The First Case Study

The majority of recordings carried out in the first case studies used the Tetris experiment, with a total of seven Tetris recording sessions being carried out. The Tetris experiment allowed for more a direct and real-time manipulation of the experimental conditions. Overall, the experiments went well, producing naturalistic clear speech. While the method of manipulation was improved in the Tetris experiments over the initial Lego experiments, it often led to demand effects (4.2.7) with some participants realising that external manipulation was taking place. Other participants approached the Tetris task in a clinical and methodological manner, keeping their voice relatively flat, giving short concise instructions and not engaging in continuous dialogue. While the overall recorded results sounded very promising the external manipulation had to be very carefully carried out so as not to arouse suspicion or create demand effects (4.2.7).

After the block-moving manipulation had taken place a number of times, participants were increasingly inclined to attribute the strange movement of the shapes to a fault or glitch in the game as opposed to one of the participants not cooperating properly. The cutting off of the audio feed only worked a few times before participants complained about the system setup not working properly, thus pulling them out of the experimental setting. One of the goals of the MIP was to minimise the disruptive nature of the experimental setup and immerse the participants in the experiment so they paid as little attention to their surroundings as possible and thus focused on the task at hand. Some of the participants were irritated by the manipulation and the hindering of their goals while others had a more positive and humorous attitude towards the hindrance, often responding with laughter. This made their responses difficult to predict: the responses to the manipulation seemed to depend somewhat on the mood and emotional state of the participants prior to doing the experiment. This is an aspect of the MIP that was beyond control, highlighting the biggest problem with MIPs: a complete lack of control over the emotional state and mood of participants

prior to the procedure. This is discussed further in section 8.3.1. The advantages and disadvantages of the Tetris MIP can be summed as follows:

Advantages:

1. It is quick and easy for participants to understand: Tetris is a well-known and popular game. Even when the rules and goals are not immediately understood they are easy to explain.
2. External manipulation is easily achieved via control of the Tetris blocks and cutting the audio feed.

Disadvantages:

1. The method of manipulation can be disrupting (especially the cutting of the audio feed)
2. Demand effects can occur (4.2.7).
3. Some participants can fail to engage in continuous dialogue, approaching the task in a clinical and methodological manner in order to succeed.
4. The responses of the participants and their attitude to the experiment were not easy to predict.

This section discussed the use of two success/failure social interaction MIPs to elicit natural emotional speech from participants. Two experiments were designed, one using Kehreins Lego based task (Kehrein 2002) and the other based on the popular Tetris game. The Lego task was used to test the recording equipment and examine methods of manipulation. The Tetris based experiment proved the more successful of the two and provided a method of real time external manipulation. However, there were disadvantages to the Tetris experimental design, the biggest being that the method of manipulation led to demand effects and detrimentally interrupted the experimental environment. Taking these findings into account, a new experimental procedure was designed to address these issues. This is detailed in the next section.

7.3 Second Case Study: Improving The MIP

This case study built upon the findings of the previous study and used modern games consoles and games instead of Tetris. The main advantage of these console systems is that a large amount of the games are usually designed with extensive multiplayer

options that are cooperative and/or competitive in nature; it was hypothesised that the inherent competitive nature of the games would elicit natural emotional responses without the need for external manipulation. It was further hypothesissed that the intrinsic entertainment value of console games would result in a largely positve group of assets and the ratings results cautiously support this hypothesis (7.4). It was not thought that the experiment would result in primary emotional states being engendered but due to the violent nature of the game this was a distinct possibility. The experimental setup is similar to the Tetris setup but a higher standard of audio equipment was used to enable a higher sampling and bit rate. Additonally, two Xbox 360 consoles were used to run a game tournament using a violent first person shooter. Computer games have previously been used to elicit emotional responses and found to be particularly suited to this task (4.2.8).

7.3.1 Experimental Setup Of The Second Case Study

The majority of participants taking part in this experiment were chosen because of their familiarity with console gaming. This ensured that participants were familiar with the overall gaming paradigm thus needing only minimal instruction in most cases. It also allowed participants to focus on achieving the in-game goals as opposed to spending the majority of time getting used to playing the game.

As discussed in section 4.2.8, computer games can allow participants to experience different emotions in a safe virtual environment, thus introducing the possibility of eliciting emotional responses that would otherwise have been difficult to engender. The use of sound booths, headphones and microphones also mirrors the setup used by Xbox 360 gamers: the console comes with a dual headphone and microphone headset and personal interaction with gamers has demonstrated that a dark comfortable setting is preferable[26]. As with the first case study, external manipulation was tested on the first group of participants using various methods; unplugging a participant's game controller, changing the time limit, giving false information regarding the amount of time left and offering a cash or material reward. It was found that the nature of the game used and the overall gaming paradigm was conducive, in many cases, to

[26] Participation and observation has shown that curtains are often closed and lights turned off in order to limit the amount of light in the room to reduce distraction and to see the screen more clearly.

inducing emotion in participants; the majority of console games have been designed to be competitive and challenging, usually with an emphasis on competitive goal achievement. The overall game play and style of most console games is therefore conducive to inducing emotional states in participants and provides an immersive virtual environment for participants. Cutting the audio feed and disconnecting controllers, as with the Tetris experiment, served mainly to disrupt the experimental conditions and break the immersive nature of the setup. Therefore no external manipulation was used in the experiments as it was hypothesised that the game provided enough internal manipulation. Figure 39 details the experimental setup, which was an elaborated version of the Lego/Tetris setup: an improved Pro-Tools HD3 system was used, allowing dialogues to be recorded 192kHz/24 bit and two Xbox 360s were used.

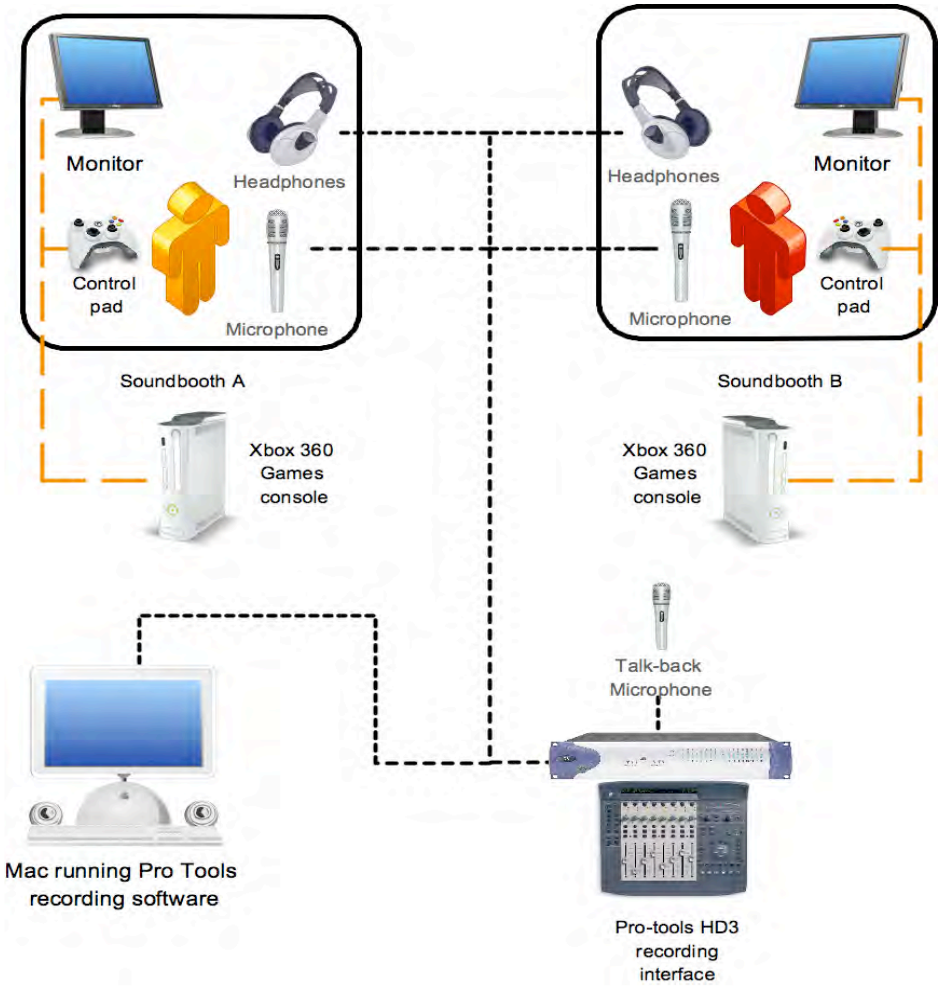


Figure 39: Experimental setup using games consoles. The setup is the same as the first case study but a higher quality Pro-Tools HD3 system was used along with Xbox games consoles.

The game used was Gears of War (GOW). GOW is a violent First Person Shooter (FPS) and involves two teams comprising of one to four people competing to survive in various small and intense multiplayer game levels. The object is to kill the opposing team before they kill you; in this case, there was only one person per team. The multiplayer levels were all small enough so that combat between the two participants was short and confrontational. Initial experiments consisted of five games per experiment with the length of games varying each time. The average length of these intial recordings was five to ten minutes long. These initial recordings included male and female gamers and non-gamers. All participants signed a consent form prior to taking part in any experiment (Appendix C).

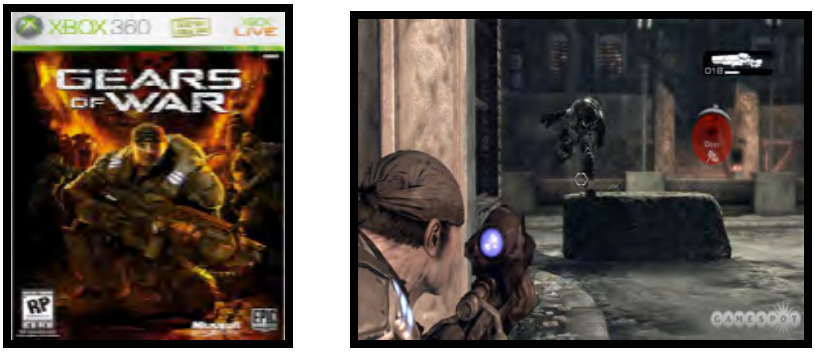


Figure 40: The Xbox 360 game, Gears of war, used in the second case study. The screenshots shows the multiplayer game play in action.

These initial recordings were used to test the game and determine the optimum game settings and in-game levels to use in the experiment. A games tournament was then created in order to attract particpants within the wider DIT Aungier/Kevin street campus. The calibration and tournament experiments attracted a total of 27 players aged between 18-28 (the majority being 18-21): twenty two of the players were male (81%) and five were female (19%). A brand new game or item of equivalent value, up to €65, was offered to the winner of the tournament. The tournament was split into three stages:

1. In the first stage each participant played a randomly selected rival. Players scored points equivalent to the number of kills they obtained in each match,

with a match consisting of five games. The two lowest scoring players in the first round were elimnated and the rest progressed to the second round.

2. In round two, each player was matched with a player they had not played in the first round and who was judged to be of equal skill. This was done to ensure the game would be challanging to both particpants in order to engender conversation over the course of the match. The two highest scoring players overall (a combination of first and second round points) played a final ten game match.

3. The final winner was the player with the most points after ten rounds. In addition to the offical tournament games, players played practice matches and cooperative games against the in-game AI. The length of each match varied with some lasting only five minutes while others lasted thirty minutes or more. In total over 200 minutes worth of audio was obtained.

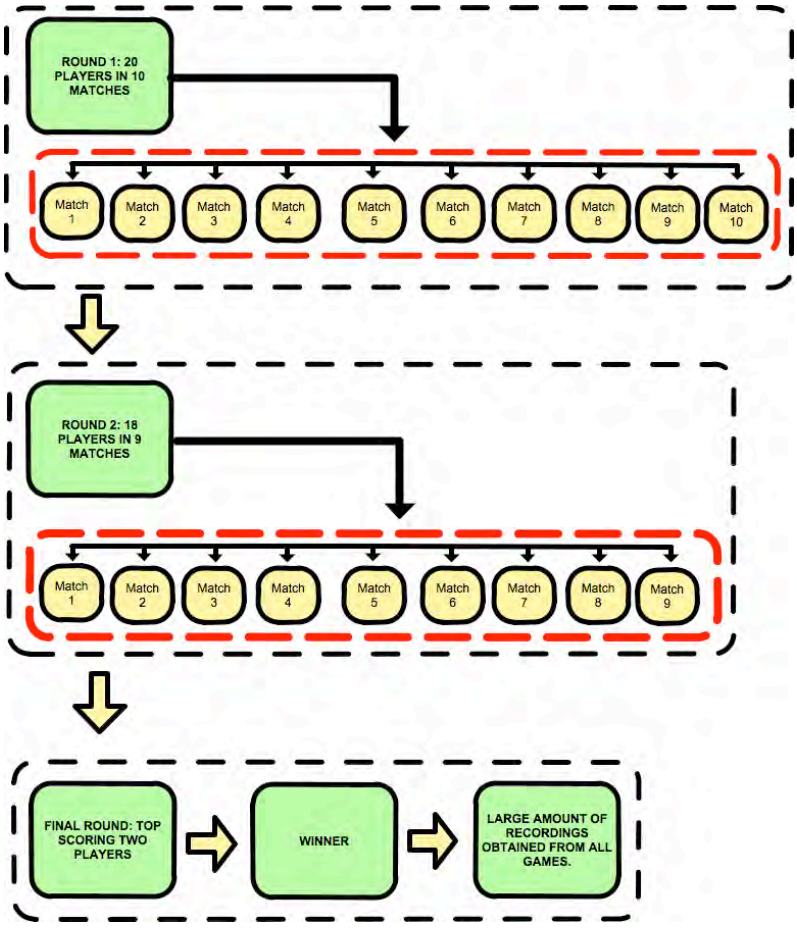


Figure 41: The organisational structure of the games tournament. A winner was decided over the course of three rounds with every game of the tournament being recorded.

7.3.2 Results Of The Game Tournament MIP

The game tournament MIP generated a large amount of speech recordings. Conversely, some of the recordings contained large amounts of relative silence as some competitors lapsed into quiet concentration. This was especially evident during non-tournament game sessions. The true purpose of the experiments was easily concealed and participants were again led to believe that human communication was the focus of the experiment. The majority of the participants were male, between the ages of 18 and 21. Only a small number of females took part. This is more than likely due to the type of game used; it is a violent FPS and appeals more to young males then females. This is a major disadvantage of using console games, especially those of a violent nature. This problem could be addressed by using a gender-neutral game or by holding separate tournaments using gender appropriate games. The experiment was more immersive than the previous Tetris experiment due to the lack of external manipulation. However, the lack of manipulation and the immersive nature of the

experiment resulted in large tracts of silence in the recordings as participants concentrated on the task at hand. Furthermore, the type of conversation tended to be limited to aspects of the game. While this was to be expected, it was hoped that conversation above and beyond this would take place. As with the first case study, different participants approached the game tournament in different ways: some took it very seriously and responded with, what could be considered, frustration, while others displayed signs of humour and laughter regardless of their performance. The advantages and disadvantages of the experiment can be surmised as thus:

Advantages:

1. Playing a competitive or cooperative computer game provides for a more immersive experimental environment.
2. Games are designed to be inherently challenging and are very well suited for use in MIP experiments.
3. Participants are willing to take place in a games tournament with a correspondingly large amount of recordings being possible.
4. The true nature of the experiment is easy to conceal with no evident demand effects.

Disadvantages:

1. The gaming tournament attracted proportionally more males than females.
2. Some participants lapsed into concentration resulting in large tracts of silence in some of the recordings.
3. Most of the conversations were related to the game environs and events. In only one case was the majority of the conversation unrelated to the game.
4. External manipulation, if used, is difficult and risks interrupting the immersive environment.
5. There are only a limited number of in-game options that can be used as a form of manipulation: time limit, score limit etc.

While this case study demonstrated that console games are a useful method of inducing emotions in participants, the type of game must be carefully considered and

possibly tailored to attract both males and females. Advanced methods of manipulation should be considered: a high level of customisation with regards in-game options would be advantageous as would timed in-game events to manipulate the progress of the participants. Consideration must also be given to the possibility that the type of game used will influence the type of emotional speech elicited. Future work beyond the current research will consider a custom designed console-based game with inbuilt emotional elicitation features (10.3.1). There was no circumstantial evidence of any primary emotional states being elicited and this was verified by the results of the asset ratings: the majority of the ratings received for the gaming assets did not lie at the extremes of the circumplex model (7.4).

7.4 Annotating The Emotional Dimensions Of The Console Gaming MIP Assets

This section discusses the emotional annotation of the assets obtained from the console gaming MIP. There were a number of issues with the obtaining of ratings for the gaming assets and this informed the method used to obtain ratings for the shipwreck MIP assets in chapter 8.

The recordings of the console games MIP were segmented into smaller assets and uploaded as per 6.3 and 6.5. A total of 624 assets were created and uploaded to the backend database. These assets served to test the upload procedure, the database integrity and the emotional dimensional annotation methodology. The batch upload process greatly decreased the amount of time necessary to upload all the assets, allowing the entire upload of the assets to be completed within a few hours. Creating the IMDI annotation data before batch uploading the assets also decreased the amount of time necessary to upload assets. Once assets had been uploaded and the integrity of the database tested (checking that files were parsed and stored correctly) the listening tool was made publicly available. Email was the primary method for advertising the tool and drawing users to it. Once the number of ratings received began to wane and slow, the listening test was stopped and the ratings were analysed.

7.4.1 Analysing The Ratings

A total of 863 ratings were received over a two-month period for the gaming assets, with only 411 of the 624 assets being rated. This equated to 65% of the assets being rated, while 35% received no rating. Removing the assets that were skipped or were unable to be rated (the Do Not Rate option on the listening test) left 397 rated assets. Considering that users were presented with blocks of ten assets with each use of the tool, dividing the total number of assets by 10 gives a rough approximation of the total number of users that took part. The total of 86 is higher than the majority of the listening groups reviewed, thus giving an average of 14 ratings per day over the two months that the listening tests took place (863/60=14.38). While a large number of ratings were received, they were not distributed evenly across all assets; the majority of assets were rated only once, with a decreasing number of assets being rated an increasing number of times. The highest number of ratings received for a singular asset was seven. Table 12 gives a break down of the number of assets within each rating group.

Times assets rated		No. of assets rated in each rating group	Percent	Cumulative Percent
1.0		167	42.1	42.1
2.0		124	31.2	73.3
3.0		58	14.6	87.9
4.0		34	8.6	96.5
5.0		7	1.8	98.2
6.0		6	1.5	99.7
7.0		1	.3	100.0
	Total	397	100.0	

Table 12: Rating count showing how many assets were rated once, twice, three times etc. Only one asset in this case received seven ratings.

The average rating per asset was 2 (863/411=2.099), with the majority of assets being rated only once (42.1%). Plotting the total number of ratings on a scatter-plot shows their distribution over the circumplex model (Figure 42).

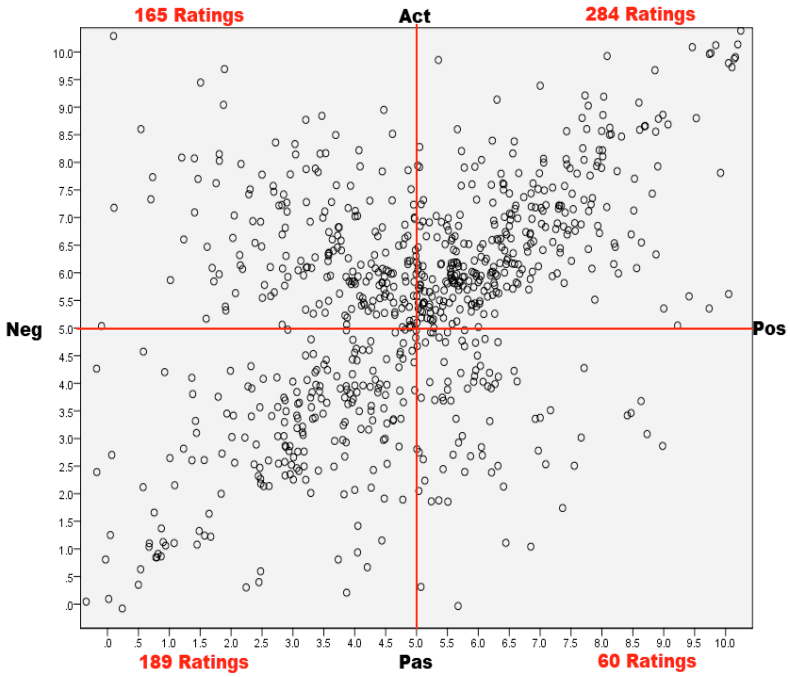


Figure 42: Jittered scatter plot of the 863 ratings received. Jittering was necessary due to the large number of similar ratings. The majority of ratings were in the active/positive quadrant.

It is evident that far too few ratings were received for any statistically relevant analysis to be carried out. While the dispersion of the ratings on the scatter plot supports the original hypothesis (that the inherent entertaining aspect of console games would result in a largely positive set of ratings) the relatively small numbers of ratings received means that this can only a tentative validation. The results of the ratings indicate that the tool performed well and that the reason for the small number of ratings is more than likely due to the large amount of assets used and the methods used to obtain ratings. Obtaining more ratings for each asset would potentially allow for a meaningful analysis of the rating results. While it was possible that a long-term rating strategy would have resulted in more ratings being obtained for the gaming assets, a more focused shipwreck MIP was devised that addressed the shortcomings of the two case studies. A smaller number of assets were derived from the shipwreck MIP and were used to obtain an increased amount of ratings (chapter 8).

7.5 Discussion

The purpose of this chapter was to investigate the use of task-based MIPs to elicit natural emotional speech from participants. The two case studies were necessary to

determine the most practical method for obtaining natural emotional speech in a high quality audio environment. Taking into consideration the discussion in chapter 4, section 4.2, it was necessary to investigate the ease at which demand effects could occur and the methods that could be employed to avoid them. While a laboratory setting is not the ideal social context in which to engender emotions, the participation of two people in each session for each task ensured that there was some level of social interaction even if it did not take place in an ideal social context. This is important as it alludes to the social communicative role of emotions (2.1.4) and incorporates the social interaction MIP (4.2.4). The Tetris experiment in the first case study was open to demand effects (4.2.7): the method of manipulation had to be carried out very carefully so as to prevent participants becoming aware of the external manipulation (7.2). A second case study was therefore developed to investigate the use of contemporary console games as a method of mood induction (7.3). This generated over 200 minutes of audio, resulting in 624 assets. While the design of the game (Gears of War) meant that it was inherently challenging, this was not enough to elicit relatively continuous emotional speech from participants (7.3.2). It was decided that a more focused MIP with a more direct set of inbuilt and external manipulators was necessary. Analysis of the rating results for the assets derived from the gaming MIP indicated that a revised method for obtaining ratings was necessary: no conclusive statistical analysis could be carried out due to the small amount of ratings received for the majority of assets (7.4.1). A revised method was necessary to obtain a significantly larger amount of ratings in order to carry out any meaningful statistical analysis.

7.6 Conclusion

This chapter examined the development of an MIP to elicit natural emotional speech in a high quality audio environment. Two case studies were carried out to test two cooperative success/failure false feedback social interaction MIPs. The first case study replicated Kehreins Lego task and also used the popular Tetris computer game (7.2). These were used to test the recording equipment and the use of external manipulation to elicit emotional responses from participants. It was determined from this that the method of manipulation could easily lead to demand effects and a second case study using a contemporary console computer game was devised (7.3). A game tournament,

with a prize offered to the winner, was held and resulted in a large number of audio recordings. There was very little external manipulation used in this case study as the game was hypothesised to inherently elicit emotional responses. While no demand effects were evident, there was not as much continuous dialogue between participants as there was during the Tetris experiment (7.3.2). The recordings from the console gaming MIP were segmented and uploaded to the backend database for rating (7.4). While a large number of ratings were received, they were unevenly distributed across the assets, with the majority of assets receiving only one or two ratings. This meant that no meaningful statistical examination could be carried out (7.5). The next chapter details a final MIP and revised rating strategy designed to address the issues raised in this chapter.

This chapter contributed to the answering of research questions: RQ1, RQ 2, RQ 3, RQ 4, RQ 5 and RQ 6.

8. A Final Task-Based MIP To Elicit Naturalistic Underlying Emotional Speech

Considering the findings of the previous chapter, a final MIP was developed to obtain more emotional speech from participants, address the problem of demand effects (4.2.7) and to refine the listening tests. While the gaming experiment was largely successful, and resulted in a large amount of assets being obtained, there were a number of shortcomings. This final MIP aimed to address these and provide for a more focused task that was attractive to both male and female participants. A superficially easier task-based co-coperative game was developed. The game was similar to a shipwreck scenario used previously by Rehm et al. in interaction experiments for the creation of a multimodal, multicultural corpus (Rehm. M., André. E. et al. 2007) and by Kousidis in experiments examining speaker convergence (Kousidis 2009). Kousidis used the shipwreck scenario to elicit natural conversation from participants and found it to be successful in engendering natural spontaneous speech from participants. His method involved presenting two participants with an imaginary shipwreck scenario and 15 items necessary for their survival. Participants were tasked with verbally agreeing on the order of importance of the items for their imagined survival (one being the most important and 15 being the least important) and it was found to be successful in eliciting spontaneous speech from the participants.

The version utilised in this chapter was adapted from Kousidis' version (which was originally adapted from Rehms version) to allow for an increased and indirect method of manipulation. As Chapter 7 illustrates, the emotional dimensional ratings of the gaming case study were, by and large, distributed over all four quadrants, with a slight concentration in the positive/active quadrant. It was hypothesised that the experiment would induce naturalistic conversational emotional speech containing underlying emotional states; it was not thought that the experimental design would induce primary or full-blown emotional responses (2.3 and 2.4). It was further hypothesised that the nature of the task would, for the most part, induce active emotional states and that the hindering of participants would induce mainly negative emotional states.

The experimental design of the MIP is first discussed, explaining the setup of the equipment and the nature of the game used (8.1.1). This is followed by an explanation of the experimental procedure and the methods of manipulation before an in-depth discussion of the results of the emotional ratings (8.2.1).

This chapter contributes to the answering of research questions: RQ 1, RQ 2, RQ 3, RQ 4, RQ 5 and RQ 6.

8.1.1 Experimental Design

The game consisted of 15 pictures of items that were needed to survive an imaginary shipwreck scenario. However, the 15 items differed from those used by Rehm and Kousidis (though in this case only 2 items differed from Kousidis' version). Clear photographs or drawings of the items were used but in some cases participants were not entirely sure what the items were but this served to foster conversation between them. The fifteen items were:

Figure 43: The 15 items used in the shipwreck MIP. All items remained static on screen throughout the experiment.

Each item image was arranged on screen in a non-hierarchical order (Figure 44):

(A) **(B)**

Figure 44: The 15 items used in the shipwreck MIP (A) and with the timer and scoring box implemented (B). The items remained static in the screen throughout to ensure participants relied on the feedback given via the score as an indication of their progress.

A basic application was created in Adobe Flex (Adobe 2009) to run the experiment and enable external manipulation of certain aspects of the experiment. A ten-minute countdown timer was added to the top of the screen along with a score box (Figure 44). Participants were told that they were the only two survivors of an imaginary shipwreck and were instructed to arrange the 15 items in order of most useful to their survival. They were told that there was a correct order for the items to be arranged in and for each item arranged correctly they would receive a point, with points being lost for items in the wrong location in the arrangement. A prize of €20 was offered to participants for achieving a score of 15 (all items arranged in the correct order). However, due to the manipulation of the experiment, this score was unobtainable.

8.1.2 Experimental Procedure

Participants were given ten minutes to rank the items: as they did this, the value in the 'Your Score' box changed. While the gaming case study (7.3) relied solely on the inherent attributes of the game being played for emotional elicitation, the results indicated a need for some form of direct manipulation. In this MIP, the scores displayed were part of an overall scoring pattern: no matter what choices were made in each experiment, the same pattern of scores was used. The score changed every time a ranking choice was made, leading the participants to believe it was in direct response to their choices. The use of a scoring pattern allowed the experimental manipulation to be standardised across all the experiments. This ensured that

differences in responses were attributable to the participants and their choices and not to changes in the method of manipulation.

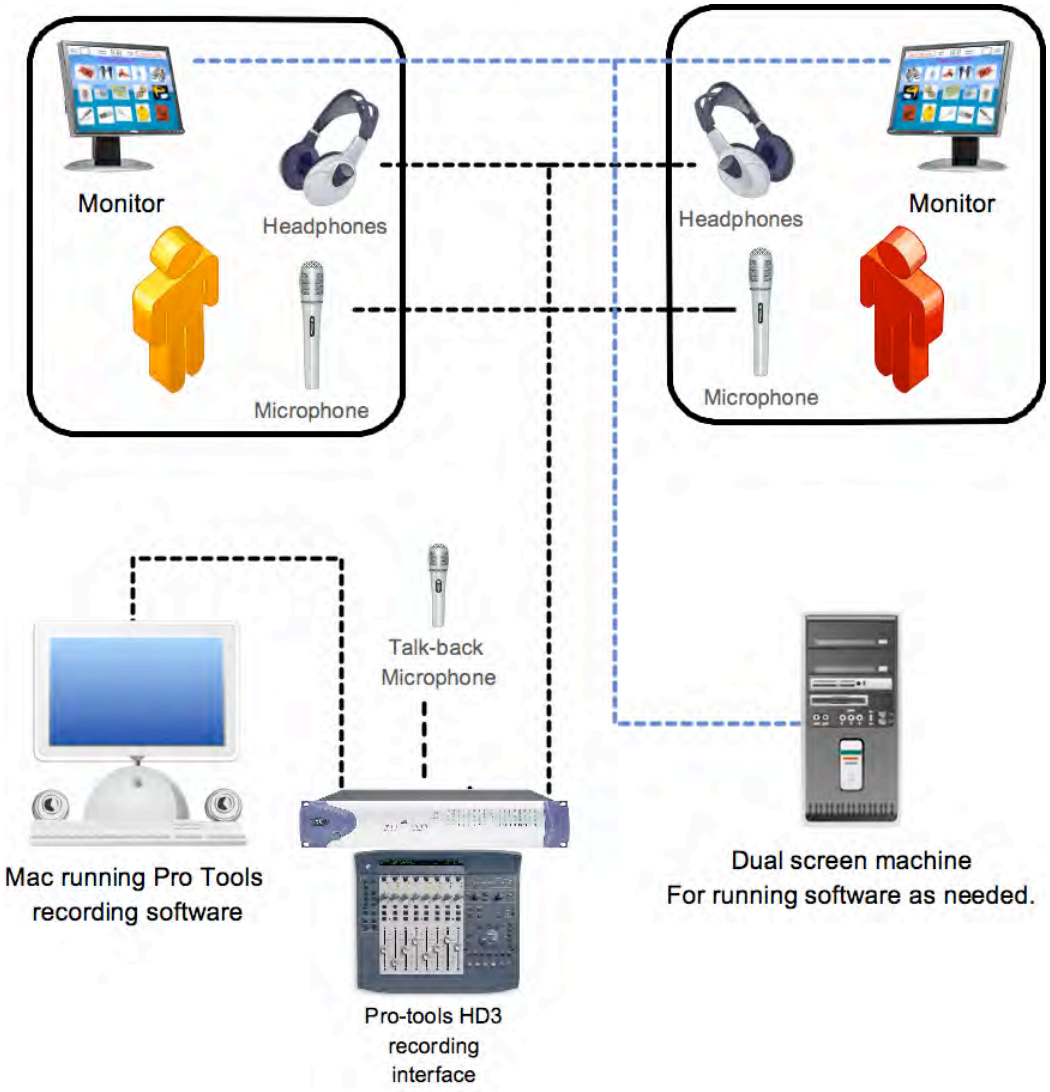


Figure 45: The experimental setup for the shipwreck MIP. The setup is the same as that of the console gaming MIP, less the Xbox consoles. An external machine is used to run the shipwreck MIP application.

As with the previous experiments, participants were seated in separate sound booths with a screen, a microphone and a pair of headphones. The recording equipment was the same as the gaming tournament case study, recording the audio at 192 KhZ/24 Bit.

The 15 items that are ranked remained static on the screen throughout the experiment; participants are told to remember the ranking choices between them. This is done for two reasons:

1. It put the participants under pressure, as they had to keep track of what order they had the items in, and had to constantly talk to each other, thus keeping a continuous dialogue going. One of the problems with the console gaming MIP was the lack of conversation and silence between participants (7.3.2) at times. Keeping the participants talking was important to ensure that as much of the ten-minute audio recording contains dialogue that could be used for emotional speech assets.

2. It made it harder to perceive the external manipulation. Since the scores were unrelated to the choices made, it might have been possible to discern this if the 15 items changed order on screen accordingly. By making the participants remember their choices, they attributed the different score to getting a ranking wrong, not remembering the order properly, or to the complex scoring system. At the very least they had to constantly communicate to verify the ranking choices and make changes in an attempt to get a higher score. Keeping the items static meant that the participants were unable to estimate how well they were doing and so had to rely on the scoring system for false-feedback (Nummenmaa and Niemi 2004) (4.2 for further discussion).

A total of 16 participants took part in the MIPs, the majority of whom were students in the DIT. As per the other two experiments, all participants signed a consent form (Appendix C) and the exact purpose and nature of the experiment was hidden in order to avoid demand effects (4.2.7). While participants were not explicitly told that the scoring system was automated, they were informed that ‘the system’ would award points based on their choices and most took this to be some form of automation. This was also done so as to avoid demand effects; awareness of an external human element rewarding marks could have led to the suspicion that scores were being manipulated. The experiment resulted in fifteen 20-minute dialogues resulting in just over 150 minutes of audio. Participants were given a few minutes prior to the start of the task to read over a set of on-screen instructions. This allowed them to understand the basic goal of the game and to allow them to acclimatise themselves with the recording environment.

8.1.3 Results Of The Shipwreck MIP

Participants responded well to the external manipulation and the time pressure of the experiment, with an increased urgency being evident in their responses, as the timer got closer to zero. This task-based cooperative MIP built upon the previous gaming based MIP, providing a more focused and easily manipulated experiment with the majority of the recordings consisting of continuous dialogue. The non-movement of the 15 items on screen as they were ranked was very successful in generating a constant dialogue between participants. The time limit and the scoring pattern, especially the negative marks towards the end of the period coupled with the decreasing time limit, appeared to successfully motivate the participants. However, as with the previous two MIP experiments, participants approached the task in a variety of different ways with some seeming to take it more seriously than others. The experimental conditions were the same in each case as was the standardised scoring pattern. The differences in the way the experiment was approached are more than likely due to individual personality traits and external factors/events prior to taking part in the experiment. While the analysis results (8.2) of this experiment demonstrated that the MIP elicited mainly active emotional speech assets, the most problematic aspect of MIPs is the lack of control over a participant's mood or emotion prior to taking part (see 7.2.4 and 8.3.1 for further discussion on this).

Since the experiment was manipulated to prevent the money being won, it was given to two participants (from the same MIP session) picked at random by a neutral party. Participants were informed, prior to the commencement of the experiment, that should no one win the money, a winner would be randomly selected; participants were also given a chocolate bar after the experiment so their participation was not entirely motivated by the prize money.

8.2 Annotating The Emotional Dimensions Of The Shipwreck MIP Assets

The recordings from the shipwreck MIP were segmented into a new instantiation of the database and uploaded as per 6.3 and 6.5. 177 assets in total were derived from the shipwreck MIP recordings. The 177 shipwreck assets attracted considerably more

ratings than the gaming assets: 3704, with all 177 assets receiving ratings, as opposed to the 863 ratings received for the gaming assets. This was most likely due to the smaller set of assets used and the increased use of Internet social media networks (described below). Dividing the total rating count by 10 gives a rough estimate of 418 as the number of raters who took part in the test: as with the gaming MIP ratings, it is the case that some raters more than likely rated more than 10 assets in one sitting. While email was the primary method of drawing users to the listening tool for the gaming assets, other avenues were explored in obtaining assets for the shipwreck MIPs. Two methods in particular attracted a large number of ratings: the use of the Twitter social media network (Twitter 2010) and the Mircoworkers website.

Twitter is a social media site that, in its own words *" …is a real-time information network powered by people all around the world that lets you share and discover what's happening now" (Twitter 2010)*. A twitter user can post short messages (140 characters in total) to their followers, people who can see what that user has posted (referred to as a tweet). Followers can then re-post (re-tweet) the original message so that their followers in turn can see it. Ideally those followers will do the same, as will their followers etc. This makes Twitter a powerful tool for the dissemination of information (Figure 46).

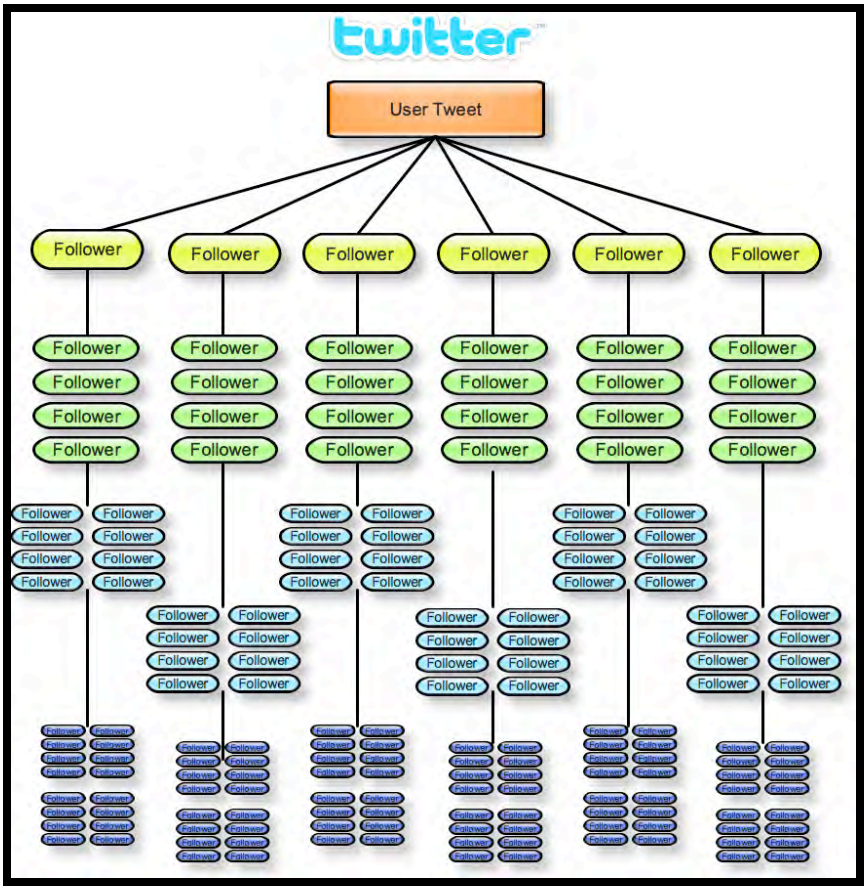


Figure 46: Conceptual flow diagram of how Twitter disseminates information: each user has a number of followers, who in turn have a number of followers and so on. Information tweeted by one user can be quickly disseminated among a large number of people.

Microworkers.com is a site that allows a large number of people (workers) to carry out short, simple tasks in return for a small fee determined by the user (employer), who requires the completion of a task or tasks (referred to on the site as a campaign) (Figure 47). An employer can advertise the task to workers around the world or restrict it to certain countries if necessary and can request a certain number of unique workers, meaning that no worker can do the task more than once. Payment for a task is awarded once an employer rates a workers work as satisfactory. Money for the payment is lodged prior to the creation of the campaign. The countries that the task can be restricted to are the USA, the UK, Australia and Canada. However the employer usually has to offer a higher fee to attract workers as a result of this restriction (Microworkers.com 2009).

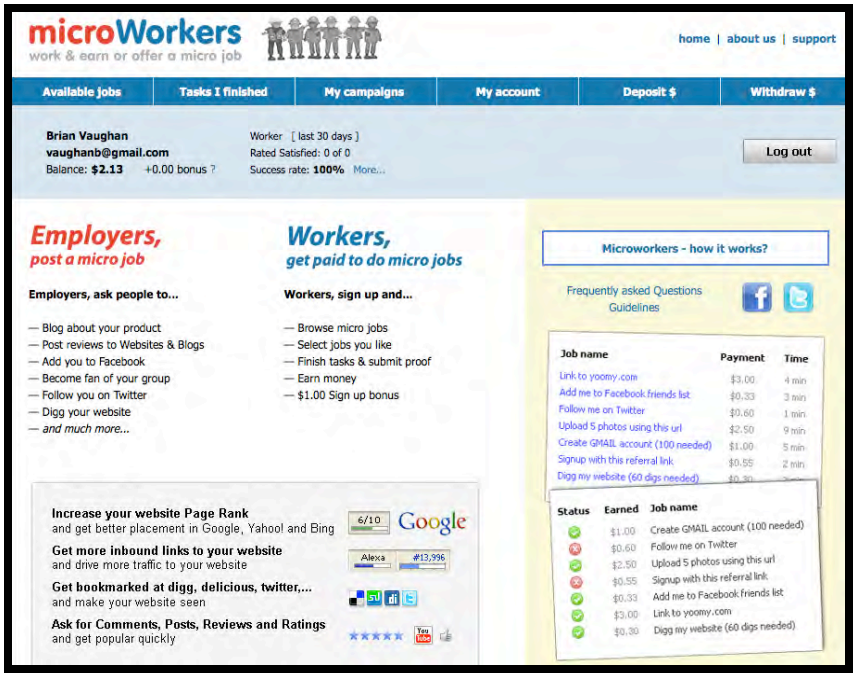


Figure 47: The Microworkers website. The site was used to obtain a large number of ratings. Raters were paid a small fee for rating a set of ten assets. The site is free to join for workers and employers.

A link to the listening tool was posted on twitter and was re-tweeted by a number of followers[27] who in turn had a large number of followers. The link was posted a number of times over the course of two months to ensure that all followers saw it. A number of relatively high profile users with a large number of followers were contacted and asked to post the link. Most obliged, thus attracting a large number of ratings. An account was created on the Microworkers site and a campaign was posted to attract workers. Workers were given a link to the listening tool instructions and were asked to provide their IP address as proof that they completed the task. The IP address of every computer used to rate assets with the listening tool was stored alongside the ratings. This enabled a rough estimation of the number of raters to be made as well as determining the country of origin of each rater using an online IP lookup service. While this is not a definitive method of determining the number of users or the country of origin, it does give an approximation[28]. A total of four

[27] Every time a tweet is re-tweeted the person responsible for the original tweet is alerted.

[28] Due to Internet Service Providers (ISPs) dynamically assigning IP addresses, it is possible for two different computers to have the same IP address at different times. Likewise, the country of origin for an IP address can often be misreported.

campaigns were carried out. The first campaign was open to workers from all countries and was used to test the Microworkers service. Workers whose first language was English were requested, with a total of 30 unique workers being requested. $0.75 was offered for rating ten assets. The owners of the site suggested that it might be better to create a new campaign restricted to workers from the US, the UK, Canada and Australia to ensure that the requirement for native English speakers. The first campaign attracted at least 270 ratings (some workers rated more than ten assets despite only being paid for rating 10) and the IP addresses of the majority of workers was verified prior to payment: three of the thirty workers did not have their IP address verified and so were marked unsatisfactory and were not paid. A second campaign was created that was restricted to workers from the US, the UK, Canada and Australia. For this campaign, $0.50 was offered for rating ten assets with 50 unique workers being requested. A total of 18 workers completed the task, which was cancelled due to its slow progress. It was determined that the amount of money offered was too low, so another campaign was created that offered $0.75 per ten ratings and looked for 30 unique workers. Progress on this campaign was improved, and as with the previous campaign, had a 100% success rate, meaning that all workers who took part had their IP address verified prior to payment. A total of 21 out of the requested 30 workers took part before the campaign was halted. Another campaign that offered $1.00 for rating ten assets and looked for 100 unique workers. This campaign proceeded quicker than the previous two restricted campaigns. At least 1000 ratings were received for this campaign, with some workers doing more than ten ratings at a time, with some giving positive feedback via the Microworkers site internal messaging system. At least 1660 ratings in total were obtained using the Microworkers website (Figure 48).

Campaign	Cost	Time	Success rate	Status	Work done	Not rated
Emotional Speech Rating For Research	$1.00	3 min		☐	100/100	-
Emotional Speech Rating	$0.75	3 min		☐	21/30	-
Emotional Speech Rating for Research	$0.50	3 min		☐	18/50	-
Emotional Speech Rating	$0.75	3 min		☐	27/30	-

Figure 48: A screenshot showing the four campaigns run on the Microworkers website. Only the last campaign was run to completion. The other campaigns were too slow and so were cancelled in favour of a new campaign with an increased amount of money offered to raters.

8.2.1 Analysing The Ratings

A total of 3785 ratings were received for the 177 assets over a two-month period. Removing the Do Not Rate ratings left 3704 ratings in total. The highest number of ratings received for any asset was 45 and the lowest was 7 (see Appendix D for the complete table) with the majority of assets receiving between 10 and 30 ratings. As with the results from the gaming MIP, dividing the total number of assets by 10 gives a rough approximation of the total number of users that took part. The total of 370 is higher than the majority of the listening groups reviewed and considerably higher than the amount of people who took part in the gaming MIP listening tests. There was a total of 62 ratings per day over the two months that the listening tests took place (3704/60=61.73). Figure 49 is a scatter plot of all the ratings received for the shipwreck MIP. What is immediately evident is that there are only a few ratings at the extremes of each dimension: this supports the original hypothesis of the shipwreck MIP that the majority of the emotional content would be underlying in nature. The majority of ratings cluster near the centre, with most lying below the 8.5 point on each scale.

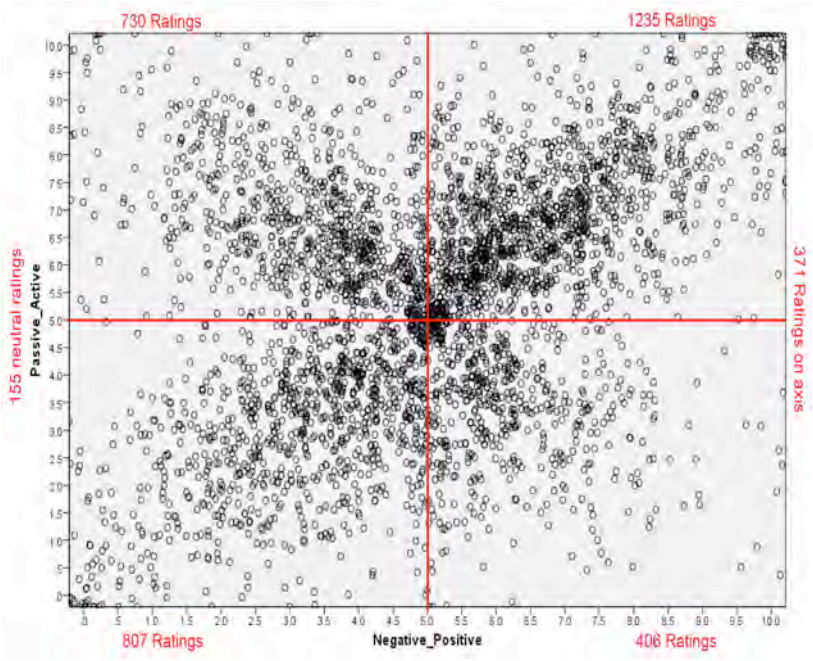


Figure 49: A jittered scatter plot of all the ratings obtained for the shipwreck MIP. The majority of ratings were in the active/positive quadrant and lay near the centre of the model and away from the edges of the dimensions.

The scatter plot gives a clearer picture of the position of the assets in the circumplex model, and shows that the majority of assets lie away from the extremes of each axis and reside in the active/positive quadrant (see Appendix E and Appendix F for tables with the median values for each asset rating group on both dimensions). Bearing in mind that the distance from the centre of the model can be considered a measurement of intensity (2.5.2), the dispersion of the assets in the scatter plot is indicative of the less intense underlying nature of their emotional content and further supports the original hypothesis regarding the shipwreck MIP.

8.2.2 Determining And Grouping According To Confidence Level

In order to determine a single rating for each asset, the central tendency (CT) and spread of the ratings received for each of the 177 assets on each dimension needed to be determined. The central tendency is the value that most represents a collection of data, and the spread is an indication of how dispersed the values are from the central tendency; a lower spread value indicates that the values are closer to the CT and less spread out. Thus a CT value with a low level of spread can be considered a relatively strong representation of the average value in a group (Peat and Barton 2005; Sharma 2005). The ratings received for each asset were treated as separate sample groups with the central tendency of each asset group being used to determine an asset's location along each dimension. With 177 asset-rating groups on each dimension, this amounted to 354 ratings groups. Examination of the sample groups revealed that they were not all of a Gaussian distribution and consequently, the mean value and standard deviation were not reliable indicators of central tendency and spread. For Gaussian distributions, the mean and standard deviation are usually used to indicate the CT and the spread of the data. In non-Gaussian distributions, the median and inter quartile range (IQR) are better indicators of CT and spread, with the IQR being less susceptible to outliers. However, the median can also be used as an indicator of CT in a Gaussian distribution, as can the IQR but only one indicator of central tendency and spread should be used when comparing distributions. The median was taken as a measure of CT for each asset-rating group and IQR as a measure of spread. With the central tendency of each asset-rating group calculated, the median values on both dimensions were plotted to position each asset within the dimensional space.

The non-Gaussian nature of a sample groups and the efficacy of using the median value and inter-quartile range to describe the CT and spread of each rating group were determined in a number of ways:

1. In a Gaussian distribution the mean and median are equal and a difference between the two values indicates a non-Gaussian distribution. Skewness and standard skewness error values indicate if a distribution is skewed to the left or right and the significance of this skew (Puri 1998; Peat and Barton 2005).
2. A Sharpio-Wilks normalcy test and Q-Q plots were used to determine dispersion. In a Q-Q plot, the presence of values on the plot deviating from the diagonal line representing expected values indicates a skewed distribution. In a Sharpio-Wilks test, a value of 0.5 indicates a non-Gaussian distribution. While not all asset ratings groups were non-Gaussian many where, and this rendered the use of the mean and standard deviation redundant as measures of CT and spread across all rating groups. (Peat and Barton 2005; Boslaugh 2008).
3. Bar charts were created for all the 354 rating groups and confirmed the findings of the other tests used.

The values obtained for these tests in PASW/SPSS (IBM 2010) indicated that a large number of the asset rating groups were skewed either left or right. This was to be expected; had the converse been true, and a Gaussian distribution was found for all rating groups, this might have indicated a problem with the MIP or the listening test methodology. To use the hypothetical example of an asset that was known, *a priori*, to be highly active: if the ratings for that asset were of a Gaussian distribution this would have the consequence that the central tendency of the group would be at the centre of the axis. Since the centre point of the axis is indicative of a more neutral emotional state, this active asset would be rated as neutral. Skewed distributions in the case of emotional dimensional ratings are expected if an MIP is designed to obtain non-neutral speech.

Median values and the IQR were obtained for all asset rating groups on each of the two dimensions; the lower the IQR the more confident the median value was as a representation of the central tendency of that group. A table with the mean, median, 25th percentile, 75th percentile and calculated IQR was created to give a more precise indication of IQR values that could then be used as a measure of confidence regarding the CT of asset groups: a low IQR value indicated a high level of confidence (Appendix E and Appendix F). The scatter-plot in Figure 50 illustrates the dispersion of the assets on the dimensional model as determined by the calculated CT value for each asset rating group..

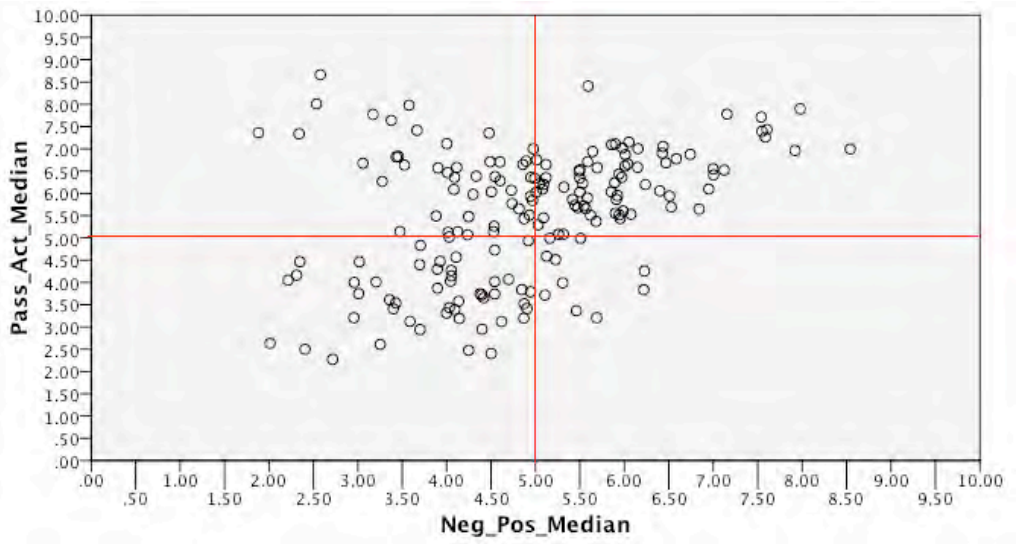


Figure 50: A jittered scatterplot of all 177 assets. This illustrates where the assets lay on the model once the central tendency of each asset rating group was calculated.Most assets lie close to the centre of the model and away from the ends of the dimensions.

Evidently, from the tables in Appendix E and Appendix F, some assets have a strong level of confidence on one dimension but a weaker level of confidence on the other. Both dimensions were examined separately to determine groups of assets with the highest level of confidence relative to each dimension. The acoustic analysis focused on the two dimensions separately in order to determine acoustic parameters relating to the level of activation and evaluation (as per RQ4 and RQ5) (chapter 9). The acoustic analysis of a group of assets with a high confidence level is potentially more meaningful than the acoustic analysis of a group with a low confidence level. An IQR value of two was taken as a cut off point: assets with an IQR of less than two were

used in a cluster analysis procedure. It was felt that an IQR of two or more was too high for any meaningful results to be obtained from the analysis of the acoustic parameters of the assets.

Clustering was used to organise the selected assets into groups of assets with the same or similar CT (Boslaugh 2008). This allowed the acoustic parameters of assets with the same/similar CT value to be analysed as a group in order to derive a set of acoustic values that were related to certain locations on each dimension. A K-means clustering procedure was used to cluster the assets into distinct groups: a number of clusters with pre-defined cluster centre values, corresponding to each of the 21 points on the dimensions (from 0-10 in 0.5 increments) were created; assets with a CT value the same/similar to the centre value of a cluster were moved in to that cluster. The aim of the process was to group all assets with similar CT values into the same cluster (Boslaugh 2008).

In a K-means clustering procedure, it is possible for members of the cluster to have values different to the centre value of the cluster; some members of the group will be closer to the centre value than others but as long as they are more similar to that cluster as opposed to other clusters, they remain within it. The difference between the value of a cluster member and the cluster centre value is described as distance and is an indication of how similar that member is to the centre value: the greater the distance from the cluster centre value, the weaker the similarity. Depending on the data, a cluster can have members that are a short distance from the cluster centre and members at a greater distance. The process of clustering the assets was carried out in a number of stages:

1. Assets with an IQR lower than two were selected.
2. 21 initial cluster centre values were specified, each one corresponding to a location on each dimensional axis, ranging from 0 to 10 in 0.5 increments.

3. A K-means cluster analysis procedure was run, using the 21 cluster centre values, to group the assets according to their CT value. In most cases only a few of the 21 clusters were filled.
4. This resulted in a number of clusters consisting of assets with the same or similar CT.

Figure 51 illustrates the process of grouping the assets according to their IQR value and then clustering them into sub-groups (based on their median values) that correspond to points on each dimension. The IQR grouping and subsequent clustering was carried out on both dimension. Appendix G and Appendix H have full details of the results of the clustering procedure.

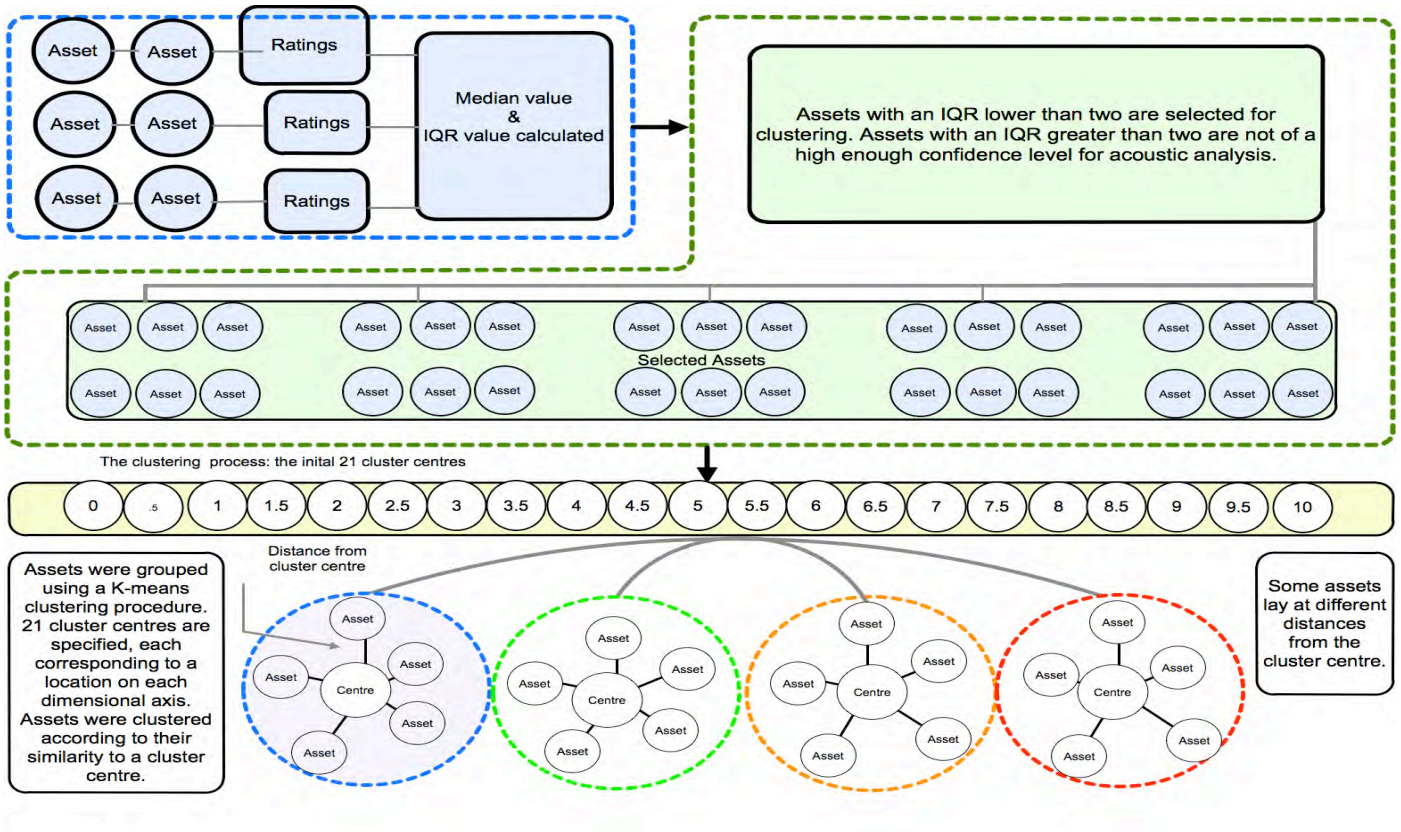


Figure 51: Flow diagram illustrating the process of grouping assets according to IQR value and using a k-means procedure to determine clusters of assets. Assets within each cluster have similar median values to each other and the cluster centre. The distance of an asset from the cluster centre is a measure of its similarity to the centre. The distance values of such assets were relatively small.

8.2.3 Cluster Results For Assets On The Activation Dimension

A total of 39 assets had an IQR value less than two and thus were grouped into cluster groups using the 21 specified possible groups as part of the k-means clustering procedure: the k-means cluster procedure described above (8.2.1) was run using the initial 21 cluster centre values listed in Table 13: these 21 values corresponded to the 21 points on each dimension. The clustering procedure resulted in nine final cluster centres (Table 14).

Initial Cluster Centres

	Cluster																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Median	0	.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10

Table 13: The initial cluster centres specified in the k-means clustering procedure.

Final Cluster Centers

	Cluster										
	7	8	9	10	11	12	13	14	15	16	17
Median	3		4		5	5.5	6	6.54	7.06	7.54	8

Table 14: The final cluster centres on the passive/active axis. There is a slight deviation from the initial cluster centre values in order to accommodate all assets.

Appendix G contains all the tables from the cluster analysis of the activation dimension. A number of assets lay at a distance from the cluster centre value: the distances were relatively small and no asset had a distance value equal to or greater than the minimum distance value between clusters. The presence of assets that lay at a distance and cluster centre values that deviated from the initial values was a result of the clustering process grouping assets according to median value similarity. These distances were small and the final centre values were rounded up to the nearest logical step on the dimension e.g.: a cluster centre value of 7.06 was rounded to 7, the closest relevant point on the activation dimension, for the acoustic analysis in the next chapter (chapter 9). Overall the procedure produced nine clusters containing assets with similar CT values that corresponded to points on the activation dimension.

8.2.4 Cluster Results For Assets On The Evaluation Dimension

A total of 52 assets had an IQR value less than two and were grouped into cluster groups using the 21 specified possible groups as part of the k-means clustering procedure: the k-means cluster procedure described above (8.2.1) was run using the initial 21 cluster centre values listed in Table 13 and resulted in 11 final clusters (Table 15).

Final Cluster Centers

	Cluster											
	5	6	7	8	9	10	11	12	13	14	15	16
Median	2.13	2.58	3	3.55	4.04	4.5	5.05	5.5	6.06	6.5	.	7.5

Table 15: The final cluster centres on the negative/positive dimension. There is a slight deviation from the initial cluster centre values in order to accommodate all assets.

As with the clustering on the activation dimension, some assets lay at a distance from the cluster centre value but no asset was at a distance that was equal to or greater than the minimum distance value between clusters. More of the final cluster centre values on the evaluation dimension deviated from the 21 initial values, but the differences were small and a logical rounding to the nearest point on the dimension was done to facilitate the acoustic analysis in the next chapter (e.g. 6.06 was rounded to 6). As with the activation dimension, assets that lay at a distance from the cluster centre, and cluster centre values that deviated from the initial values, were a result of the clustering process grouping assets according to median value similarity. Appendix H contains tables with all the results of the cluster analysis for the evaluation. Overall the procedure produced 11 clusters containing assets with similar CT values that corresponded to points on the evaluation dimension.

The next section discusses the use of MIPs and large-scale listening tests.

8.3 Guidelines For Obtaining And Evaluating High Quality Natural Emotional Speech.

This section is a discussion of the use of MIPs to elicit emotional responses from participants and the use of online-based large scale listening tests to rate their emotional dimensions. Various factors need to be considered in designing MIPs and this chapter discusses these elements in light of the findings in chapter 7 and sections 8.1.3 and 8.2.1.

8.3.1 The Use Of MIPs

Both the gaming case study and the final shipwreck MIP were successful in eliciting emotional responses from participants. While the gaming MIP relied on the game itself to provide the emotional eliciting elements, the shipwreck game relied on internal (the countdown timer) and external manipulation (the scoring pattern). Ideally a game could be specifically designed with manipulation elements coded into it, a combination of the two approaches. These elements could take the form of timed in-game events or in-game events triggered internally after a certain event has occurred or externally using a simple key press. This would ensure that the eliciting events take place within the immersive game environments. Future work will consider the development of such a game (10.3.1).

The gaming MIP scatter plot of ratings revealed that most of the ratings and assets resided in the active/positive quadrant. Likewise with the scatter plot for the shipwreck MIP; however, there were also a large amount of ratings and assets in the active/negative quadrant. While the eliciting elements in the shipwreck MIP were intended to frustrate the participants and hinder the attainment of the goal (to get a maximum score of 15), some participants reacted positively to the manipulation (the scoring pattern, countdown timer and the static on-screen items). This may have been due to a number of reasons that are discussed in the following paragraphs.

The participants in the shipwreck experiment were friends in order to ensure a high level of conversation between the two. The Tetris and gaming case studies suggested that participants that were friends were better at initiating a relatively continuous

dialogue between them. In the gaming experiment there were large tracts of silence as some participants lapsed into quiet concentration. This happened most between participants who had only met during the gaming tournament. A subjective overview of the recordings indicated that when two participants who described themselves as friends took part, there was a marked increase in the amount of conversation, with participants often discussing topics beyond the game being played. There is an argument to be made for using participants who describe themselves as friends: they are comfortable in each other's company and would not feel as socially (2.1.4) constrained in their expression of emotion as two strangers would. There was a marked difference in the amount of conversation between participants in the gaming and shipwreck MIPs. This was due to different experimental designs and conditions but the fact that the participants in each shipwreck MIP experiment session were friends should be considered as an important factor.

In the case of the shipwreck MIP, the countdown timer and scoring pattern hindered the participants in completing the task. It was hypothesised that this would lead to largely negative emotional responses from participants. However the results obtained indicate that this is not the case. While a large amount of ratings and assets were placed in the active/negative quadrant, the majority of assets and ratings are situated in the active/positive quadrant. The monetary incentive offered for the completion of the task was supposed to be an added impetus to the participants. It is possible that this amount, €20, was too small. A larger amount of €100 or so would have been a more desirable prize. Having said that, it is potentially unethical to offer a large sum of money for a task that is knowingly being manipulated so as to prevent its completion. As detailed (8.1.2) the €20 prize offered in the shipwreck MIP was given to one of the participants selected at random. All participants were told, when taking part, that if no one completed the task and won the prize it would be given to a randomly selected participant.

While some of the problems with the MIP could possibly be addressed with an altered experimental design, there are aspects to the MIP that are uncontrollable and unknowable. Although MIPs take place in a controlled laboratory setting, there is no

control over the external environment outside the lab. While an MIP can be used to induce emotion in a laboratory setting, there is no control and little knowledge about participants' mood and/or emotional state prior to entering the lab or taking part in the MIP. A series of questions and measurements can be asked prior to the experiment, this increases the chances of demand effects colouring the results; the true purpose of the MIPs were hidden from participants and any assessment of participants mood or emotional state prior the experiment would have potentially revealed the true purpose of the experiment. Even with these uncontrollable elements, it is argued that a combined task based false feedback social interaction MIP offers the best method of eliciting natural emotional responses from participants: no experimental process can account for every variable; the most that can be hoped for is a high degree of control of certain aspects. Total control over all aspects of the MIP experimental process, while not necessarily possible, may not be advantageous if the desired outcome is natural underlying emotional speech. As Chapter 4 discusses, the problem with simulated sources of emotional speech is their artificially controlled nature and the more control exerted over an MIP, the more one risks hampering the naturalness of the conversation.

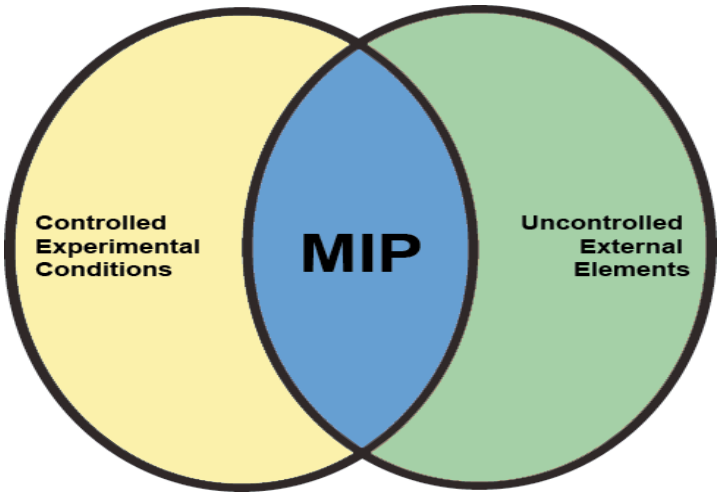


Figure 52: Venn diagram of the juxtaposition of the controllable and the uncontrollable elements in an MIP. A certain loss of control is necessary to obtain truly natural emotional speech.

8.3.2 Large-scale Listening Tests

As the results of the shipwreck MIP listening tests demonstrate (8.2.1), by using social networking a large number of ratings can be obtained. By doing a reverse trace on the IP addresses of each rater, stored each time the listening tool is used, the

country of origin of each rater can be determined[29]. This revealed that the majority of raters came from the US, followed by Ireland, the EU, and Great Britain. Table 16 gives a detailed break down.

IE	US	EU	GB	ID	IN	CA	PH	CN	DE	AU	RO	LV	CH	BD	AR
1185	1362	487	263	89	70	169	64	50	42	82	30	20	10	40	10

Table 16: Breakdown of the number of ratings received from each country based on IP address. IE=Ireland, US=United States, EU=European Union, GB=Great Britain, ID=Indonesia, CA=Canada, PH=Philippines, CN=Switzerland, DE=Germany, AU=Australia, RO=Romania, LV=Latvia, CH=China, BD=Bangladesh, AR=Argentina.

The distribution between the 16 countries was due to the methods used to obtain assets. Twitter and the Microworkers site are worldwide sites and are not necessarily restricted to specific countries. While certain countries can be stipulated in setting up a campaign on the Microworkers site, this relies on the member workers having submitted correct location information. With Twitter, once something has been sent out on the service there is very little control over who sees the tweet and who accesses it. As it is, the instructions given to the workers on the Microworkers site specifically asked for native English speakers and the Twitter request was sent out to Twitter users for whom it was known that English was their first language. However, there is an element of trust in using both methods to obtain ratings. While it is not the purpose of this research to ascertain if there are differences in the ratings given to assets by listeners from different countries, one must keep in mind that this could be an influencing factor in the ratings received. Future work in this area will consider the ratings given to assets by raters from different countries to investigate this aspect further (10.3.3).

Aside from the consideration of possible cultural differences, using online listening tests is largely dependent on listeners carrying out the test in the correct manner: free from distraction and with an understanding of how to correctly use the tool; especially the use of the two separate scales to rate the emotional content of the speech. In a

[29] While IP addresses are a useful method of determining the location of raters, it is not fool proof. Owing to the nature of the World Wide Web and the allocation of IP addresses, the country of origin for each IP is not a definite.

laboratory setting listeners can be observed carrying out the ratings and can be fully informed by the researcher if they are unclear about the use of the listening tool. This is obviously not possible in using an online listening tool; again an element of trust enters the equation, namely that listeners can clearly hear the audio and understand how to correctly use the listening tool. The instructions prior to the tool were simple and clear to best explain the purpose and use of the tool while also avoiding leading the listeners by using subjective emotional terms in the explanation. Despite the degree of control afforded by laboratory testing, it is argued that online listening tests offer a distinct advantage over small-scale laboratory based tests by allowing for the collection of a larger data set from a wider population.

Listening tests can also be used as a method of evaluating the MIP used. In the case of the gaming MIP, the majority of the ratings and assets lay in the active/positive quadrant, thus tentatively confirming the hypothesis that a console game would elicit generally positive emotional responses. As discussed, console games are designed to be entertaining and it would be expected that this would be reflected in the type of ratings received. Considering that the game used was of the violent-action genre, one would also expect a level of activation. However, the fact that not all assets in the gaming experiment were rated and that only a small number of ratings were achieved for the majority of the rated assets means that this can not be considered a concrete validation of the MIP. However, the large ratings received from the shipwreck MIP also confirm this hypothesis. While the shipwreck MIP is not identical to the gaming MIP, it is a game-based task, and one would expect a certain amount of positive ratings from it. It was also hypothesised that the external manipulation would result in a large amount of negative ratings due to participants being hindered in their task. While a number of ratings resided in the active/negative quadrant, the majority of assets were in the active/positive quadrant. Considering this, the results of the listening test would suggest that the shipwreck MIP is generally suited to the elicitation of active/positive emotional states. Conversely, one must consider the uncontrollable elements already discussed as an influencing factor in the type of emotional states induced. It would seem then that the type of underlying emotional response elicited by an MIP depends on a number of factors. The unpredictability of the responses of participants may be due to the fact that the MIP was designed to obtain less intense underlying emotional states. The elicitation of primary emotional

states would most likely require a stronger level of directed emotional elicitation, as in the case of Tolmkitt and Scherer who used images of skin disease and injured human bodies to induce emotional responses in participants (Tolkmitt and Scherer 1986), and therefore should result in a more consistent set of emotional responses.

8.4 Discussion

The imaginary shipwreck scenario was successful in producing high quality audio recordings of natural underlying emotion. Inbuilt manipulation was achieved through a very prominent countdown timer that was constantly present on screen, while external manipulation was achieved through the promise of a financial reward for correctly ordering the items and through the use of a set scoring pattern entered by an external researcher; regardless of the choices made by all participants the scoring outcome was the same. This resulted in 177 assets in total. As with the two case studies (chapter 7), the response of participants to the experiment varied.

While the small number of ratings obtained from the gaming MIP (7.4.1) was insufficient to be used for in-depth analysis, the large number of ratings for the shipwreck MIP could be grouped and were clustered into distinct groups (8.2.2). The revised rating strategy attracted a large number of raters and ratings: 370 raters and 3704 ratings for the shipwreck MIP assets compared to 86 raters and 863 ratings for the gaming MIP assets. This was due to the increased use of social networks in obtaining ratings (8.2). The 863 ratings received for the gaming MIP assets were widely dispersed across the assets, with only one asset being rated seven times; the majority were rated only once or twice. This was too small an amount of ratings per asset for any meaningful grouping or statistical analysis of the rating results: no determination of the central tendency for each asset could be reliably determined. In contrast, the large number of ratings received for the shipwreck MIP enabled the central tendency of each asset-rating group to be determined and thus grouped according to their IQR value that served as a measure of confidence. This allowed assets with a strong confidence rating (an IQR of less than 2) to be grouped together for clustering using a k-means clustering procedure (8.2.1). The k-means clustering procedure resulted in groups of assets that were clustered together due to their similar median values, which were in turn similar to the centre value of the cluster they were

in. The centre values of each cluster were identical or similar to one of the 21 points on each dimension (0-10 in 0.5 increments) and thus assets in the clusters could be placed at one of the 21 points on each dimension. Clustering the asset ratings without any prior confidence grouping would have resulted in a large number of clusters with a relatively weak level of confidence, containing (as they inevitably would) assets whose median was highly representative of their central tendency and assets whose median was not highly representative of their central tendency. The grouping and clustering method adopted created a small set of asset clusters with a high confidence rating, thus lending more weight to the findings of the acoustic analysis in chapter 9. While this reduced the amount of assets that were to be subjected to acoustic analysis, the argument is made that this small group of assets with a strong confidence level provided for a more robust analysis group. If an asset has a median value that is not robustly representative of its central tendency, the results of any acoustic analysis would possibly be weaker.

The distance values in each cluster group can also be considered another indication of confidence: the more asset rating group median values that lie at a distance from the cluster centre to which they belong, the less representative the cluster centre value is of the asset rating groups within it. However, the distance values were relatively small; any values of 0.5 or more would have put that asset into another cluster. Incidents of assets lying at a distance from the centre value were due to a number of assets in the cluster having median values that deviated slightly from the 21 initially specified values. Likewise with cluster centre values that deviated from the 21 initially specified values: the deviated final cluster centre values were necessary to adequately group the assets in the cluster. While MIPs are useful for eliciting emotional responses, the response of participants is not always predictable (8.3.1). Participants in the shipwreck MIP reacted differently to the internal and external manipulation. This may be due to uncontrollable elements prior to the participant taking part in the MIP and the fact that underlying as opposed to primary emotional states were being elicited. A certain loss of control is necessary in experimental procedures designed to elicit natural underlying emotional speech: too much control and manipulation can result in demand effects and hamper the induction of truly natural emotional speech. Likewise with the use of large-scale web-based listening tests: a certain loss of control is necessary to obtain a large number of ratings.

8.5 Conclusion

This chapter examined the creation of a final MIP to elicit natural underlying emotional speech, based on the findings of the previous two case studies (7). The results of the rating of the assets from this final MIP were also considered, demonstrating that the revised rating strategy, which used web based social networking, vastly increased the amount of ratings obtained (8.2). The twitter social network and the Microworkers were used and it was from these sources that most of the ratings were derived. Analysis of the ratings received for each asset revealed that they were of a non-Gaussian distribution and therefore the median and Inter Quartile Range were used as indicators of central tendency and spread (8.2.1). This enabled the IQR to be utilised as a measure of confidence regarding the median value of each asset: a low IQR value meant that the median value was a reliable representation of the central tendency of each asset-rating group (8.2.2). Once the IQR values were calculated, all assets with an IQR value of less than two on each axis were subjected to a k-means clustering procedure. The cluster analysis specified 21 initial cluster values that corresponded to the 21 points on each dimension, with final cluster centre values deviating slightly from the original 21 in only a few cases (8.2.3 and 8.2.4). Grouping and clustering the assets in this manner created a small set of statistically robust assets that could then be acoustically analysed. Although the groups of assets were small, it was argued that the results of the acoustic analysis would be more robust as a result. A number of recommendations were made regarding the use of MIPs and large-scale listening tests, determining that there are certain aspects of MIPs that are beyond the control of the researcher and as a result the outcome of MIPs is not always predictable (8.3). The use of large scale listening tests concluded that their use entails a certain loss of control over who takes part and that cultural differences between raters would need to be considered in future work.

This chapter contributed to the answering of research questions: RQ1, RQ 2, RQ 3, RQ 4, RQ 5 and RQ 6.

9. Analysis Of The Acoustic Parameters Of Natural Underlying Emotional Speech Assets

The purpose of this chapter is to carry out a preliminary examination of the acoustic parameters of the corpus speech assets with similar ratings from the shipwreck MIP. As the results of chapter 8 demonstrated, the majority of the ratings and related assets sit close together on each axis and are underlying rather than primary in nature (2.4 and 2.5). Assets with an IQR of less than two on each axis were selected for acoustic analysis due to their higher confidence rating. These assets were then grouped according to their rating on each axis and had a number of their acoustic parameters measured: pitch mean, intensity mean, pitch range, intensity range, pitch contour, jitter and shimmer values, spectral slope, spectral centre of gravity and speech rate. A Spearman's rank correlation procedure was used to determine if there was a correlation between each parameter and the activation and evaluation dimensions. Initially the acoustic parameters of male and female speakers were considered separately, however due to the small number of male speakers in the group of assets used in the acoustic analysis, it was decided not to analyse the parameters relative to gender. This did not adversely affect the results of the correlation procedures.

The analysis procedure is first detailed, followed by a reporting of the analysis results for each axis in relation to each of the acoustic parameters measured. A discussion section examines the results in relation to the findings of the literature review in chapter 3 before the conclusion of the chapter.

This chapter contributes to the answering of research questions: RQ1, RQ 2, RQ 3, and RQ 6.

9.1 Acoustic Analysis Procedure

The PRAAT audio software was used to perform the acoustic analysis. As detailed in 5.2.1, PRAAT provides a comprehensive set of audio analysis procedures, including the ability to write custom scripts to carry out these procedures. PRAAT scripts were used in the development of LinguaTag and are a powerful way of utilising PRAATs

functionality for custom acoustic analysis and the batch analysis of a large number of audio files. While LinguaTag can be used for the measuring of certain acoustic parameters, the extent of its analysis capabilities was insufficient for the level of audio analysis needed. A PRAAT script was developed to carry out such a batch analysis, providing a comprehensive set of acoustic parameter values (Appendix I). Alongside the custom script written to carry out the analysis, an automatic syllable nuclei detection script was used. This script was written by Nivja H. de Jong and Ton Wempe who found that it worked well when compared to manual speech rate calculation and was well suited for comparing the speech rate of different speakers (De Jong and Wempe 2009) (Appendix J). Considering the findings of chapter 3 regarding the acoustic parameters of emotional speech, a number of acoustic parameter values were obtained for each of the target assets on each dimension of the model: pitch mean, intensity mean, pitch range, intensity range, pitch contour, intensity contour, jitter and shimmer values, spectral slope, spectral centre of gravity and speech rate.

9.1.1 Pitch Analysis

The pitch was calculated in Hertz and semitones in PRAAT using the custom written script that incorporated a two-pass pitch detection procedure (Appendix I). Daniel Hirst of the PRAAT user group recommended this procedure (PRAAT-Users-Group 2010). The two-pass pitch detection procedure adjusts the pitch analysis range relative to each speaker and avoids octave errors, where octaves of the F0 are misreported as base F0 values. The use of semitone values is more useful for comparing pitch values across groups and speakers and is indicative of the degree of change rather than the absolute value. Consider the case of a male and female speaker increasing their pitch values in an utterance. The male speakers pitch increases from 150Hz (86.745 st)[30] to 300Hz (98.745st), an increase of 150Hz or 12 semitones (98.745st-86.745st). The female speakers pitch increase from 300Hz (98.745st) to 600Hz (110.745st), an increase of 300Hz or 12 semitones (110.745st-98.745st). The comparison of the semitone values in this case is more meaningful than the linear Hz, as the semitone values tell us more about the change in pitch relevant to each speaker: regardless of their respective pitch ranges, both speakers increased their pitch by the same ratio.

[30] The formula: $12x\log_2(Hz)$ is used to covert Hz to semitones (st).(Patel 2006)

Reporting semitone pitch values allows pitch changes across dimensional rating groups to be meaningfully compared. Pitch contours were computed and drawn within PRAAT and displayed the pitch changes of each speaker over the course of an utterance (3.2.1).

9.1.2 Intensity Analysis

Prior to segmentation and upload (6.5) all assets had their intensity normalised using Pro-Tools audio software to account for speaker dependent variables such as distance from the microphone and orientation in relation to it. Normalisation rescales an audio file such that all amplitude peaks are brought up or down to a common value (usually 0dB) with the rest of the amplitude points in the file being brought up by the same relative scaling factor (Owren and Bachorowski 2007). Hence all the assets had the same peak amplitude value, thus aiding comparison between rating groups. The intensity median and range was obtained for all assets.

9.1.3 Voice Quality: Jitter And Shimmer Analysis

While it is beyond the remit of this research to determine and define the acoustic correlates of voice quality descriptors (0), jitter and shimmer measurements were obtained due to their apparent relation to certain voice quality descriptors such as "roughness" and "hoarseness" in speech (0).

9.1.4 Voice Quality: Spectral Energy Distribution

Spectral energy has been suggested as a parameter related to certain voice quality descriptors in that increased high frequency energy is often described as being ‘harsh’ sounding and having a ‘bright’ or ‘brilliant’ quality (0). It has also been found to be positively correlated to the activation dimension of emotional dimensional models (3.6). Scherer and Juslin and Laukka also found spectral energy distribution to be related to certain emotional states (3.6). Values for the spectral slope and centre of gravity were obtained as a measure of the spectral energy distribution within each asset. The spectral slope is a measure of the rate of decrease in amplitude of the frequency content of the speech signal and is measured in decibels (dB) per octave. Slope values are given as minus dB values and the bigger the negative value, the greater the amount of high frequency roll-off there is; this is often described as affecting the timbre of a sound as discussed in (0) (Baken and Orlikoff 2000) (Figure

53). A centre of gravity value was also obtained for each asset. The centre of gravity, or first spectral moment, is a measure of spectral balance and indicates the frequency region were the most frequency energy is concentrated: a higher value indicates increased high frequency energy (Harrington and Cassidy 1999; Spackman, Brown et al. 2009).

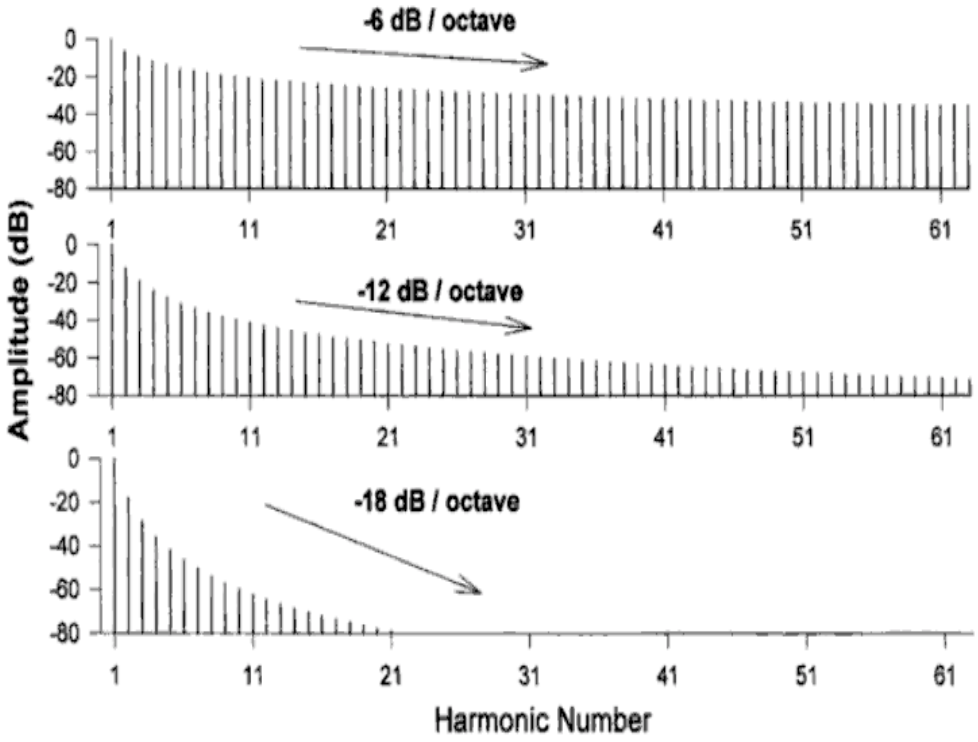


Figure 53: A diagram illustrating the change in intensity of a frequency spectrum from loud (top) to soft (bottom). The higher the negative dB value, the greater the amount of high frequency roll-off there is. Conversely, the lower the negative dB value, the greater the amount of high-frequency energy there is in the signal. Taken from (Baken and Orlikoff 2000).

9.1.5 Speech Rate Calculation

Considering the discussion of speech rate in 3.3.4, speech rate and not articulation rate was measured. Speech rate was calculated by dividing the duration of each asset by the number of syllables within to get a syllable per second speech rate value. The aforementioned script written by De Jong and Wempe (De Jong and Wempe 2009) was used to calculate the number of syllables in each asset.

9.1.6 Determining Correlation

SPSS (IBM 2010) was used to carry out a two-tailed Spearman's rank correlation procedure to determine if there was a correlation between each acoustic parameter and the passive/active dimension and/or the negative/positive dimension. The formula for Spearman's rank correlation is:

$$r_s = 1 - \frac{\left(6\sum d^2\right)}{n\left(n^2 - 1\right)}$$

Equation 4: Equation for calculating Spearman's Rank Correlation.

Where r_s has the range $-1 \leq r_s \leq 1$ with $r_s = 1$ implying a perfect positive correlation, $r_s = -1$ implying a perfect negative correlation and $r_s = 0$ implying no correlation at all. In the above formula, d is the difference in rank between two numbers and n is the number of pairs of data being compared. The procedure works by ranking two sets of data, from highest to lowest and assigns a value based on rank with 1 being the highest rank; the difference in rank is squared (d^2) along with the n value and the correlation is determined using the above formula. The Spearman's rank correlation is used for non-Gaussian distributed data and is often used as the non-parametric version of Pearson's correlation coefficient (Puri 1998; Boslaugh 2008). Two different correlation significance levels of 5% (0.05) and 1% (0.01) were returned for different acoustic parameters: a 0.05 significance level indicates that there is a 5% chance that random samples of data will show no correlation, while the 0.01% significance level indicates that there is only a 1% chance of it occurring (Bryman and Cramer 2004). Median values of every acoustic parameter for each dimensional group were calculated. The median rather than the mean was used due to its resilience to outliers and robust indication of the central tendency of data with a non-Gaussian distribution [31]. Appendix K and Appendix L have the Spearman's report for each acoustic variable for each dimension along with trend line scatter plots.

[31] As with 8.2.2, the distribution of the data for each acoustic parameter was generally not of a Gaussian distribution.

9.2 Correlation Of Acoustic Parameters On The Activation Dimension

Appendix K contains the tables for the Spearman's correlation procedures and relevant graphs for each acoustic variable. The number of assets in each cluster group is given in Table 17.

Cluster Group Value	No. of assets in each group	Percent	Cumulative Percent
3.00	1	2.6	2.6
4.00	2	5.1	7.7
5.00	1	2.6	10.3
5.50	1	2.6	12.8
6.00	4	10.3	23.1
6.50	12	30.8	53.8
7.00	9	23.1	76.9
7.50	6	15.4	92.3
8.00	3	7.7	100.0
Total	39	100.0	

Table 17: The number of assets in each cluster group on the activation dimension. The cluster group with the centre value of 6.5 had the most assets.

9.2.1 Pitch

Spearman's rank correlation revealed significant correlation between pitch median (Hz) and the level of activation. There was a positive correlation of 0.750 between pitch and activation, significant to 0.05% with a sample size of 9: r=0.750, n=9, p=0.020. The results were similar for the semitone pitch values (r=0.750, n=9, p=0.020) with a significance of 0.05. There was also a strong positive correlation between the level of activation and pitch range, (r=0.850, n=9, p=0.004). This would appear to support the findings of the literature review in chapter 3 where average pitch and pitch range were found to positively correlate with increased activation level. However the semitone pitch range values indicated no correlation with the activation dimension (r=0.300, n=9, p=0.433). This is likely due to the fact that the semitone pitch values are a better measure of the relative change in pitch range between

activation categories, and thus the relative change is not correlated to an increase in activation level, as are Hz pitch range values.

9.2.2 Pitch Contour

Manual analysis of the pitch contours for each cluster group [32] ascertained that no conclusive trend could be ascertained for the pitch contours in each activation group. The overall trend of the contours was also inconclusive with only two rating groups having contours that demonstrated the same overall trend: contours in group four all had a general downward trend while group 7 had eight (out of nine) contours with a downward trend.

9.2.3 Intensity

There was a strong negative correlation for intensity with the activation dimensions at the 5% level: r=-0.700, n=9, p=0.036. Conversely, intensity range had a positive correlation with activation level at the 5% level: r=0.667, n=9, p=0.050. The reason for this inverse relationship between intensity range and intensity median in relation to the activation dimension was due the fact that the median minimum intensity decreased with increased activation, while the maximum intensity remained stable. This led to an increased intensity range and a lower median intensity. A dual scatter plot for median minimum intensity and median maximum intensity (Figure 54) clearly indicates this to be the case as did the Spearman's rank correlation results (see Appendix K) there was a strong negative correlation between median minimum intensity and activation level, r=-0.750, n=9, p=0.020 but none for median maximum intensity, r=0.33, n=9, p=0.381. However, the low n value of some of the groups here means that no definitive conclusion can be made.

[32] Analysis of activation groups with n=1 were not examined.

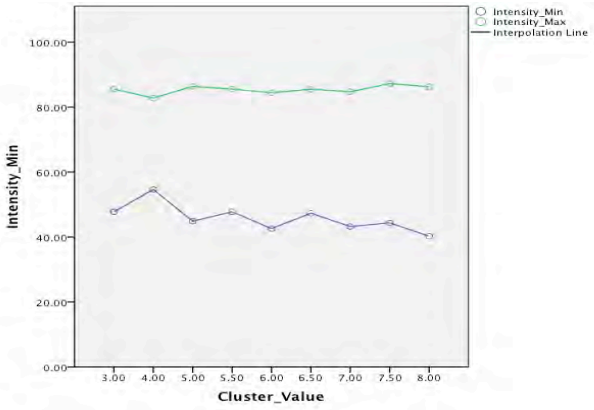


Figure 54: Dual trend scatter plot of the median maximum intensity and the median minimum intensity. The maximum intensity remains stable while the minimum intensity decreases as the level of activation increases. This accounts for the positive correlation of the intensity range, and the negative correlation of the median intensity, with the activation dimension.

9.2.4 Jitter And Shimmer

Jitter and shimmer demonstrated no correlation with the activation dimension (r= -0.350, n=9, p=0.356 for jitter and r=0.033, n=9, p=0.932 for shimmer). No results for either were considered in the literature review other than research indicating that they were responsible for rough and hoarse voice quality descriptors.

9.2.5 Spectral Energy

There was a strong positive correlation between activation and spectral slope, and between activation and centre of gravity at the 1% significance level. Previous parameters were correlated at the 5% level of significance but results of the correlation procedure in SPSS returned values at the 1% significance level for these parameters, indicating a greater level of significance. For slope the correlation was, r=0.817, n=9, p=0.007 and for centre of gravity it was, r=0.900, n=9, p=0.001. Some researchers found increased high-frequency energy to be positively correlated with the activation dimension of a dimensional model (3.6). Considering the literature review in chapter 3, this suggests that increased activation leads to a 'brighter' sounding speech.

9.2.6 Speech Rate

There was no correlation between speech rate and level of activation, r=0.533, n=9, p=0.139. This is contrary to the findings in chapter 3 in relation to speech rate and the

level of activation which found speech rate increased with the level of activation. This is something that needs to be examined further in conjunction with articulation rate calculations as future work 10.3.5.

9.3 Correlation Of Acoustic Parameters On The Evaluation Dimension

Appendix L contains the tables for the Spearman's rank correlation procedures and relevant graphs for each acpustic parameter. The number of assets in each cluster group is given in Table 18.

Cluster_Group_Value

Cluster Group Value	No. of assets in each group	Percent	Cumulative Percent
2	2	3.8	3.8
2.5	3	5.8	9.6
3	2	3.8	13.5
3.5	5	9.6	23.1
4	6	11.5	34.6
4.5	6	11.5	46.2
5	11	21.2	67.3
5.5	4	7.7	75.0
6	8	15.4	90.4
6.5	2	3.8	94.2
7.5	3	5.8	100.0
Total	52	100.0	

Table 18: The number of assets in each cluster group in the on the evaluation dimension. The cluster group with the centre value 5 had the most assets in it.

9.3.1 Pitch

The Spearman's rank correlation procedure revealed no correlation between median pitch and evaluation (r=-0.155, n=11, p=0.650) and between median pitch range and evaluation (r=-0.482, n=11, p=0.133). Likewise for median pitch and pitch range

measured in semitones: r=-0.145, n=11, p=0.670 for median semitone pitch and r=-0.400, n=11, p=0.223 for median semitone pitch range.

9.3.2 Pitch Contour

Manual analysis of the pitch contours for each cluster group [33] revealed that no conclusive trend could be ascertained for the pitch contours in each evaluation group: numerous groups had contours with a high degree of variability and with a low degree of variability. Only four rating groups had contours with the same level of variability: 3.5, 4, 4,5 and 7.5. Similarly with the overall trend of the contours, only a few rating groups had contours that demonstrated the same overall trend: rating group 3, which only contained two contours in it and rating group 6.5, which also only contained two contours in it. In light of the findings here and in 9.2.2, consideration should be given to examining methods of systematic pitch contour analysis, perhaps using the ToBI system or

9.3.3 Intensity

As with pitch, there was no correlation between median intensity and median intensity range with r=0.155, n=11, p=0.650 for intensity and r=0.173, n=11, p=0.612 for intensity range. This substantiates the findings of the literature review in chapter 3 which found that correlating acoustic parameters with the evaluation aspect of emotional states is difficult and inconclusive.

9.3.4 Jitter And Shimmer

As with the activation axis, there was no correlation between jitter and evaluation and between shimmer and evaluation, with r=-0.055, n=11, p=0.873 for jitter and r=-0.91, n=11, p=0.790 for shimmer.

9.3.5 Spectral Energy

Unlike the correlation results for spectral energy on the activation axis, there was no correlation found between spectral slope and the evaluation axis, r=0.227, n=11, p=0.502 and between centre of gravity and the evaluation axis, r=-0.191, n=11, p=0.574. While some of Schröder, Cowie et al. findings suggested a correlation

[33] Analysis of evaluation groups with n=1 were not examined.

between high frequency energy and the evaluation dimension, the results here indicate otherwise (Schröder, Cowie et al. 2001) (3.6).

9.3.6 Speech Rate

As with the correlation results for the activation axis, no correlation was found between speech rate and evaluation (r=0.255, n=11, p=0.450). As with most of the results for the acoustic parameters in relation to the evaluation axis, this also corroborates the findings of the literature review. Schröder, Cowie et al. suggested that positive evaluation was correlated with a faster speaking rate, as with spectral energy, but the results here also suggest otherwise (Schröder, Cowie et al. 2001) (3.6).

9.4 Discussion

The results of the acoustic analysis and correlation suggest that certain acoustic parameters of underlying emotional states are related to the activation dimension of the circumplex model but there appears to be no correlation for any of the acoustic parameters and the evaluation dimension. Table 19 details the results of the correlation analysis for each axis with the results discussed in detail below.

	Activation	Evaluation
Median Pitch	0.750* (Hz & st)	-0.155 (Hz) –0.145 (st)
Pitch Range	0.850* (Hz) 0.300 (st)	0.482 (Hz) –0.400 (st)
Pitch contour	Inconclusive	Inconclusive
Median Intensity	-0.700*	0.155
Intensity Range	0.667*	0.173
Spectral Slope	0.817**	0.227
Centre of Gravity	0.900**	-0.191
Jitter	-0.350	-0.055
Shimmer	0.033	0.91
Speech Rate	0.533	0.255

* Correlation is significant at the 0.05 level
** Correlation is significant at the 0.01 level

Table 19: A summary of the findings regarding the acoustic parameters of the analysed assets. There is a correlation between the activation dimension and median pitch, pitch range, median intensity, intensity range, spectral slope and the centre of gravity (high-lighted in green). There is no correlation between the evaluation dimension and any of the acoustic parameters.

9.4.1 Correlation Of Pitch With The Activation And Evaluation Dimensions

On the activation axis, median pitch has a strong correlation with level of activation (r=0.750). However, while the median semitone pitch value is correlated in the same way as the median Hertz value, this is not the case for the median pitch range: the Hertz value demonstrates a correlation while the semitone range does not. This is due to the fact that semitone range values are a better indicator of the ratio of change between the minimum and maximum pitch values. The pitch range in both cases is calculated by subtracting the minimum pitch value from the maximum pitch in both Hertz and semitones, rather than calculating the range in Hertz and then converting it to semitones or vice versa. The semitone values are calculated in the PRAAT script and give the semitone pitch range as the difference between the minimum semitone pitch and the maximum semitone pitch. As demonstrated in 9.1.1, although two sets of pitch ranges appear to be different when the Hertz values are compared, the semitone value is a better indication of the relative difference. In the example given in 9.1.1, the male's range is 150Hz and the female range is 300Hz, but both have a range of 12 semitones relative to their minimum and maximum pitch values. The semintone pitch range values are a better value to use in determining if there is a significant difference between cluster groups. On the activation dimension, the relative semitone difference between cluster groups is therefore not as significant as the linear Hertz range difference between cluster groups. In contrast, there was no correlation for Hertz and semitone pitch and pitch range on the evaluation dimension. There was no correlation for the median pitch or median pitch range in relation to the level of evaluation. The findings regarding pitch contour on both dimensions were inconclusive, with only a few groups exhibiting any clear trend.

9.4.2 Correlation Of Intensity With The Activation And Evaluation Dimensions

Median intensity demonstrated a strong negative correlation with the activation dimension (r=-0.700) while intensity range demonstrated a strong positive correlation (r=0.667). This was due to a negative correlation between the level of activation and median minimum intensity: the median minimum intensity decreased as activation increased while the median maximum intensity remained stable, thus lowering the

median intensity value and increasing the median intensity range. This is contrary to the findings of the literature review in chapter 3, where increased intensity and intensity range appear to have a positive correlation with activation level. There was no correlation for median intensity and median intensity range in relation to the level of evaluation, in keeping with the findings of chapter 3. Future work will be carried out to examine minimum and maximum intensity values on a wider set of assets (10.3.5).

9.4.3 Correlation Of Spectral Energy With The Activation And Evaluation Dimensions

There was a positive correlation for the two measures of spectral energy: spectral slope and centre of gravity both indicated increased high frequency energy with a corresponding increase in activation. The correlation for these two parameters was the highest of all the parameters analysed in this chapter and correlation for both was significant at the 1% level. This corroborates some of the findings in chapter 3. Angry speech was found to have increased high frequency energy and would generally be considered an active emotional state, while sad speech was found to have decreased high frequency energy, and would generally be considered a more passive emotional state (Pereira 2000; Schröder, Cowie et al. 2001) (3.6). However, the consensus regarding high frequency energy in relation to other emotional states is not as robust; there was no consensus for fearful, happy or disgusted speech. Schröder, Cowie et al. found high frequency energy to be correlated with the activation dimension of an emotional model (Schröder, Cowie et al. 2001)(3.6). Moreover, high frequency energy is not considered by all researchers and is not one of the core acoustic parameters usually investigated in the literature in relation to emotional speech. There was no correlation for spectral slope or centre of gravity and the evaluation dimension. Future work will examine spectral energy in relation to the activation and evaluation dimension on a wider set of assets (10.3.5).

9.4.4 Correlation Of Jitter, Shimmer And Speech Rate With The Activation And Evaluation Dimensions

No correlation was found for jitter, shimmer or speech rate on either dimension. The findings of chapter 3 indicated that speech rate was positively correlated with certain emotional states. This relationship does not appear hold true for underlying emotional

states. Speech rate also appeared to be positively correlated with pitch and intensity: when pitch and intensity were increased so was speech rate. This may be due to speech rate being positively correlated with the activation dimension along with pitch and intensity and hence would appear to be positively correlated with either parameter. However, the fact that pitch and intensity were positively correlated with the activation dimension of underlying emotional states while speech rate was not, indicates that there is no correlation between speech rate and either of these parameters in this context.

While a direct comparison of the results in this chapter with the findings of chapter 3 is difficult due to the use of different methodological and theoretical approaches, some of the results in this chapter support those in chapter 3. The findings for median pitch, pitch range and intensity range are in agreement with the findings in chapter 3 regarding their positive relationship to the level of emotional activation. Spectral energy also demonstrates a positive correlation with activation, something that is tentatively suggested by the findings of the literature review in chapter 3. More work is needed in this area as it is not as widely considered in the literature; the results for pitch median and range, intensity median and spectral energy indicate that there is a strong correlation with the activation dimension. Conversely, speech rate was not correlated with the activation dimension while median intensity was negatively correlated. These findings contradict the findings in chapter 3 and may be due to the different methodologies used in this research compared to the majority of the literature reviewed in chapter 3. The approach taken here is staunchly dimensional, while the majority of studies in chapter 3 utilise prototypical primary emotional categories thus making a direct comparison of results difficult. Only a small number of studies have considered dimensional models and acoustic parameters (3.6). While the development of dimensional models often involved the mapping of emotion terms or categories into the dimensional space, no conclusive one-to-one mapping is evident in the literature. The fact that no correlation, positive or negative, was found for any of the parameters in relation to the evaluation dimension substantiates the consensus in chapter 3.

The lack of correlation for parameters in relation to the evaluation dimension suggests that a wider set of acoustic parameters needs to be considered, specifically in relation to this dimension, both at the primary emotional level and the underlying emotional level. The fact that the findings of this chapter support some of the findings of the literature review in chapter 3, especially regarding pitch and spectral energy, suggests that the two-dimensional model was successful in capturing salient aspects of emotional experience. The findings in chapter 3 suggested a correlation between certain acoustic parameters and the activation aspect of emotional experience: the results of the analysis in this chapter also suggest this to be the case. As argued in 2.5.3, a dimensional model of activation and evaluation represents the activation and appraisal aspect of emotional experience. The fact that a general pattern of correlation was found for certain acoustic parameters and the activation dimension of the two-dimensional model used (with regard to underlying emotional speech), supports the hypothesis that certain acoustic parameters are positively related to the activation aspect of emotional experience. The fact that underlying emotional states were the focus of the analysis in this chapter may be a factor in the contradictory results for median intensity and speech rate in relation to the activation dimension. While a positive relationship is evident for pitch in relation to full-blown and underlying emotions, this does not mean that a positive relationship will also be found for intensity. The characteristic mild intensity of underlying emotions (2.3) may also be a factor in the lack of a positive relationship for median intensity: variations in intensity may be too small to be of significance in the case of underlying emotional states. This may mean that intensity is not a useful parameter to measure to consider in relation to underlying emotional states. This fact, coupled with the lack of correlation for any acoustic parameter in relation to the evaluation dimension, for both primary and underlying emotional states, suggests that alternative acoustic parameters need to be investigated in relation to both dimensions.

9.5 Conclusions

This chapter examined ten acoustic parameters in relation to the activation and evaluation dimension of a circumplex model. Positive correlations were found for median pitch, spectral slope and centre of gravity in relation to the activation dimension (9.2). No conclusive correlation was found for median pitch range on the

same dimension: the median semitone pitch range showed no correlation, while the median Hertz pitch range showed a significant correlation (9.2.1). Median intensity was negatively correlated with the activation dimension while intensity range was found to be positively correlated with the activation dimension (9.2.3). This was due to the negative correlation of the median minimum pitch value with the activation dimension, having the effect of lowering the overall median intensity value while increasing the overall median intensity range. No acoustic parameters were positively or negatively correlated with the evaluation dimension (9.3). Overall, the findings regarding some of the acoustic parameters in relation to activation dimension and the lack of correlation for any parameter on the evaluation dimension substantiate some of the findings of the literature review in chapter 3 despite methodological and theoretical differences.

This chapter contributed to the answering of research questions: RQ1, RQ 2, RQ 3, and RQ 6.

10. Conclusions

10.1 Summary Of Work

This thesis documents an investigation into the creation and analysis of a natural underlying emotional speech corpus. A review of existing perspectives and definitions of emotion was first undertaken (chapter 2), investigating the four main psychological perspectives on emotion (2.1.), the neurological aspects of emotion (2.2.1) and the physiological aspects of emotion (2.2.2). Theories of full-blown and underlying emotions (2.3) and primary and secondary emotions (2.4) were examined, as were dimensional models of emotion (2.5). This chapter of the review suggested that although there is a consensus regarding a small set of primary emotions, there is no definitive list of emotions. It also argued that underlying emotional states are an important part of the communicative process and therefore their investigation is more relevant. Emotional dimensional models were determined to be a more objective method for describing emotional states than emotion categories.

The next chapter of the review (chapter 3) documented existing work regarding the acoustic correlates of certain primary emotional states. The definition of prosody across three disciplines was considered (3.1), demonstrating that a more focused definition of prosody allowed for a more rigidly defined set of acoustic parameters to be utilised. A meta-analysis of the literature regarding the acoustic correlates of five primary emotional states was undertaken (3.3). It was demonstrated that the majority of the literature found a positive correlation for pitch, intensity and speech rate in relation to the activation aspect of emotional states. Although the correlation is for primary emotional categories, the findings did suggest that these acoustic parameters were important in the expression of emotional speech and therefore merited investigation in relation to underlying emotional speech (3.6).

Existing emotional speech corpora and the sources of their emotional speech were then considered (4.1). It was demonstrated that simulated emotion is the most used method for obtaining emotional speech and that simulated emotional sources and broadcast sources are unreliable sources of natural emotional speech (4.1.1 and 4.1.2).

Mood Induction Procedures were then considered as methods of obtaining natural emotional speech (4.2), determining that they are ideally suited for obtaining emotional speech in a high quality-recording environment (4.2.9). Audio quality was next considered, examining current audio standards and arguing for the use of a high sample and bit rate in order to future-proof and archive the recordings (4.3).

Methods of annotating emotional speech assets were then reviewed. A three-tiered annotation approach was proposed in order to carry out metadata annotation (5.1), acoustic annotation (5.2) and emotional annotation (5.3). The advantages of using a large sample size to rate the emotional content of speech was then investigated, making the argument that larger sample sizes are advantageous compared to smaller sample sizes and can result in a more robust set of statistical results (5.3). A three-tiered approach is necessary to ensure the comprehensive annotation of corpora assets to ensure their reusability and usefulness and to enable comparison between different methodological approaches (5.4).

As a result of this review, several areas of research were examined that led to the formulation of a number of research questions by which the scope of this thesis could be defined:

10.1.1 RQ 1: Is a two-dimensional model adequate to capture some salient aspects of natural underlying emotional speech?

This research question was formulated in chapter 2 after examining the various perspectives and definitions of emotion in the literature (2.1-2.4). Commentators disagree on a definitive list of primary or full-blown emotional states, and it was determined that underlying emotional states are an important aspect of human communication, with the majority of everyday emotional experiences arguably being underlying in nature (2.3). It was established that the use of dimensional emotional models was a more objective method of describing emotional states (2.5). A review was carried out of existing emotional dimensional models, demonstrating that the activation and evaluation dimensions are employed in the majority of models (2.5). While some commentators have suggested more than two dimensions are necessary, two-dimensional models have been successfully used by numerous researchers (2.5.2) and were adopted in this research.

This research question was answered in chapters 7, 8 and 9. Chapter 7 carried out two case studies to examine the use of MIPs as a method of obtaining natural underlying emotional speech. Results from this chapter led to the development of an improved MIP and rating strategy in chapter 8. Results of the statistical analysis suggested that the speech obtained contained underlying emotional states as opposed to primary or full-blown emotional states as was the purpose of the MIP (8.2.1). The analysis results of chapter 9 corroborated some of those in chapter 3, with differences in results for some of the acoustic parameters examined possibly being due to differing methodologies, and the different emotional approach taken: underlying emotions as opposed to primary emotions were examined. Based on the conclusions of chapters 7 and 8 and the results of chapter 9 it is concluded that a two-dimensional model is adequate to capture some salient aspects of natural underlying emotional speech but an expanded range of parameters needs to be considered in relation to two-dimensional models of emotion.

10.1.2 RQ 2: Can certain acoustic parameters of natural underlying emotional speech be correlated with the activation dimension of a two-dimensional circumplex model?

This research question was formulated in chapter 3 after a review of the literature regarding the acoustic parameters of five emotional states (3.3). A strong consensus was evident across the literature regarding five emotional states (anger, fear, happiness, sadness and disgust) and certain acoustic parameters (pitch mean, pitch range, intensity mean, intensity range, speech rate and pitch contour) (3.3.1 - 3.3.4). It was also determined that pitch mean, pitch range, intensity mean, intensity range and speech rate were positively correlated with the activation aspect of emotional speech: as activation level increased so did the values for these acoustic parameters. There was some consensus regarding high frequency energy between two of the commentators.

This research question was answered in chapter 6, 7, 8 and 9. Chapter 6 devised a listening tool to obtain a large amount of ratings for emotional speech assets and implemented a three-tiered annotation approach within a backend database. Chapters

7 and 8 devised methods of obtaining natural underlying emotional speech for rating using the listening tool as well as a statistical method for analysing the ratings obtained. Chapter 9 carried out the analysis of the acoustic parameters of the obtained speech with the highest confidence level, based on Interquartile Range (IQR) values. The analysis carried out in chapter 9 indicated a positive correlation for median pitch, pitch range, intensity range and high frequency energy with the activation dimension of the model (9.2). While contradictory results were found for intensity median (a negative correlation) and speech rate (no correlation), this may be due to methodological differences: the fact that underlying emotional states were being analysed is one major difference. The relative low-intensity of underlying emotional states is also possibly a factor (9.4). Considering the findings of chapter 3 and chapter 9, it can be concluded that certain acoustic parameters of natural underlying emotional speech can be correlated with the activation dimension of a two-dimensional circumplex model.

10.1.3 RQ 3: Can certain acoustic parameters of natural underlying emotional speech be correlated with the evaluation dimension of a two-dimensional circumplex model?

As with research question 2, this research question was formulated in chapter 3 after a review of the literature regarding the acoustic parameters of five emotional states (3.3). The findings of the literature discussed in chapter 3 indicated that it was difficult to correlate acoustic parameters with the evaluation aspect of emotional states (3.6). The results obtained in chapter 9 support this finding: no correlation was found for any acoustic parameter in relation to the evaluation dimension of the emotional model. Considering the findings of chapter 3 and chapter 9, it can be concluded that the acoustic parameters examined in chapter 9 cannot be correlated with the evaluation dimension of a two-dimensional circumplex model. A wider set of acoustic parameters would need to be considered to determine if it is these parameters alone that cannot be correlated or if a wider set of parameters cannot be correlated with the evaluation dimension (10.3.5).

10.1.4 RQ 4: Can a practical MIP based experiment be designed and used to elicit natural underlying emotional speech from participants in a high quality audio environment?

This research question was formulated in chapter 4 after a review of the literature regarding existing emotional speech corpora; the methods used to obtain emotional speech; the use of MIPs as a method of obtaining natural emotional speech and the use of high quality audio standards for recording natural emotional speech using an MIP. The use of simulated and broadcast sources of emotional speech was determined to be unreliable with regard to the naturalness of their emotional content (4.1). MIPs were determined to be the best option for obtaining natural emotional speech as long as the correct MIPs were used (4.2). Some MIPs have the same problem of authenticity that simulated and broadcast sources do (4.2.1 - 4.2.3 and 4.2.5). A cooperative social task-based success/failure MIP was determined to be the best approach to obtaining natural underlying emotional speech (4.2.9) as well as avoiding demand effects (4.2.7). The review of current audio standards concluded that high definition audio is the current professional standard and is the de-facto consumer standard in some regards (4.3.3 and 4.4). For future proofing and archiving purposes, speech should be recorded at the highest possible standard and converted to a lower level of quality when practically necessary (4.4.2).

This research question was answered in chapters 7 and 8. Chapter 7 carried out case studies to examine the use of MIPs to obtain natural emotional speech. The experiments in chapter 7 and 8 used soundproof booths to provide clean and clear audio. The first case study in chapter 7 used a recording quality of 96kHz/24bit (7.2.1), and the second case study used a higher quality 192kHz/24bit (7.3.1). In both cases audio recordings were clean, clear and noise free. The final MIP in chapter 8 also used a recording quality of 192kHz/24bit with a clean, clear and noise free signal being the result (8.1.3). While the first case study in chapter 7 was subject to demand effects, the second case study was not. Likewise the final shipwreck MIP was not subjected to demand effects and, based on the analysis of the ratings (8.2.1 and 8.4), it was successful in obtaining natural underlying emotional speech. Based on the conclusions of chapter 7 regarding the gaming MIP and the conclusions of chapter 8,

a practical MIP based experiment can be designed and used to elicit natural underlying emotional speech from participants in a high quality audio environment.

10.1.5 RQ 5: What are the practical considerations of annotating an emotional speech corpus?

This research question was formulated in chapter 5 after purposing a three-tiered approach to the annotation of emotional speech corpora. This approach annotates corpora in three different ways: metadata annotation, acoustic annotation and emotional dimensional annotation (chapter 5). Existing metadata schemas were examined and it was determined that the IMDI schema was the most suited to the metadata annotation of speech assets (5.1.1). The acoustic annotation of assets was then considered. The LinguaTag software was examined in relation to the annotation of the acoustic parameters of speech assets. While the range of parameters that LinguaTag uses is limited, its method of storing the analyses results allows them to be easily parsed into a backend database (5.2.2). The PRAAT speech analysis software can be used to analyse the speech assets for an expanded range of acoustic parameters. However its results cannot be parsed into a backend database in the same way as LinguaTags, but it does allow a comprehensive acoustic analysis to be carried out (5.2.1). Emotional annotation was then considered, determining that large rating groups are preferable to small rating groups, with the majority of studies utilising small groups of raters to determine the emotional dimensions of speech (where dimensional models are used) or to categorise the emotional states of speech (usually into primary emotional categories) (5.3).

This research question was answered in chapters 6, 7 and 8. Chapter 7 and 8 investigated the use of MIPs to obtain natural emotional speech assets for the use in the creation of natural underlying emotional speech corpus. In chapter 6 a persistent backend database was determined to be the ideal method of bringing the three forms of annotation together in a logical and structured manner, in relation to the assets obtained in chapters 7 and 8. Doing so allowed the corpus to be connected to, and utilised via, current web based technologies (6.4.2 and 6.4.4). The IMDI schema was instantiated and adapted as part of the backend database that was created using the Ruby-on-Rails architecture. The IMDI schema was implemented in such a way as to

allow IMDI metadata to be created prior to or as part of the asset upload procedure (6.2.5). The output of the LinguaTag acoustic analysis could then be parsed into the database to sit along side the metadata in order to test the veracity of parsing and storing acoustic annotation (6.3). Once assets had been uploaded and parsed, large-scale web-based emotional ratings could be carried out. This was achieved through the use of a dimensional rating tool, created in Adobe Flex as a browser based interface. This allowed the rating tool to be accessed from any computer with a web browser and Internet connection (6.4.2). An interactive corpus visualisation tool was also created to allow for an intuitive and aesthetically pleasing method of viewing the assets and their acoustic parameters within the corpus (6.4.4). Chapter 6 concluded that the full and complete annotation of emotional speech assets and their storage in a logical and structured backend database was important to addresses issue of data sparcity; to ensure that assets could be easily accessed and utilised via web based tools; to enable the widespread distribution of the corpus via the use of personalised logins; and to ensure that the large amount of work that was used to record the assets and develop the corpus resulted in a comprehensively annotated corpus that could be utilised in a wide variety of research.

Considering these conclusions, there are a number of practical considerations in annotating an emotional speech corpus:

- A metadata annotation schema should not only provide for the annotation of demographic data but should also allow the methods used to obtain the emotional speech to be comprehensively documented. This allows for a comparison of different emotional elicitation methods as well as enabling researchers external to the original data collection to understand the methods used.

- The annotation of the acoustic parameters of the speech should be stored in a file format that can be parsed into a backend database. This reduces the amount of time necessary to fully annotate the acoustic aspects of emotional speech. It also allows the acoustic data to be utilised and queried in the same manner as the metadata.

- The emotional annotation of a speech corpus can only take place once the assets are structured in a backend database that enables them to be accessed and utilised using web based technologies.

- A persistent backend database is necessary to store the different levels of annotation and must ensure that all forms of annotation can be queried and utilised as fully as possible. This avoids data redundancy and addresses issues of data-sparcity within the field of emotional speech research. It also allows a corpus to be easily distributed to other researchers, via personalised logins, when necessary.

- Extra cost and work is needed to implement a three-tiered annotation schema within a backend database. However, doing so has the potential to reduce long-term economic and administration costs via improved access to the data as well as providing data relevant to researchers in other fields (speech linguistics, anthropology etc).

10.1.6 RQ 6: What are the advantages and limitations of using a large population size in rating the emotional dimensions of speech assets?

This research question was first formulated in chapter 5 after examining the use of listening groups in the literature (5.3). It was determined that the majority of listening groups use only a small number of listeners to determine the emotional categories or dimensions in the speech being examined. The use of small listener groups is problematic in that they are more susceptible to chance findings, best-fit errors, and they may not result in a true representation of the emotional content of the speech being examined.

This research question was answered in chapters 6, 7, 8 and 9. Chapter 6 implemented the three-tiered annotation schema proposed in chapter 5, and chapters 7 and 8 used MIPs to obtain natural underlying emotional speech which was segmented, annotated and stored using the methods implemented in chapter 6. The assets were then subjected to statistical analysis to determine the dimensional rating for each asset. For the gaming MIP in chapter 7, not enough ratings were received for the assets (624 in

total) for any meaningful statistical analysis to be carried out. A revised method for obtaining emotional ratings was devised in chapter 8, along with a smaller set of assets (177 in total), that made use of social networking sites to obtain ratings. A significantly increased amount of ratings was obtained for all assets as a result. This allowed for a meaningful statistical analysis to be carried out which resulted in a smaller set of assets, derived from the original 177 assets, with a strong confidence rating to be used for acoustic analysis.

A further discussion on the subject of MIPs and large-scale rating tests concluded that large-scale listening tests needed to be web-based to reach a large number of participants. However, in doing so a certain amount of control over the listening test participants and the conditions in which the listening tests are carried out is lost; there is no way of knowing if the listening tests were carried out in an ideal acoustic environment, what the cultural background of the participants is or if the testing software ran properly on the participants machine. Both these elements have the potential to influence the ratings of the assets. Though steps can be taken to control access to online listening tests, this does not fully address the loss of control over certain aspects of the listening tests. Intelligent use of a wider set of social networking sites was proposed as a possible solution. Once a confident set of assets had been determined they were acoustically analysed in chapter 9 with a correlation between certain parameters and the activation dimension being evident.Considering these conclusions, a number of points regarding the use of large populations sizes to rate the emotional dimensions of speech assets can be made:

Advantages:

- A large amount of geographically dispersed people can be reached, thus potentially attracting a large amount of ratings.
- Using a large sample size ensures that results from statistical and acoustic analysis are more robust and less likely to open to the same errors inherent in using a small sample size.
- Subsequent acoustic analysis will be more robust and relevant as a result.

Limitations

- There is a loss of control over the participants taking part and the acoustic environment within which they listen to the assets.
- Using a large-scale web based listening tests requires a structured and robust backend database to feed assets to the listening tool and to store the ratings. This increases the time and cost in implementing a large-scale web based listening test.
- A lot of time is needed to obtain a large amount of ratings for all assets being rated. As the gaming MIP demonstrated, a large amount of ratings does not necessarily mean that the ratings will be distributed evenly among all assets. The shipwreck MIP addressed this by using a smaller set of assets. An increased amount of assets is preferable for acoustic analyses to provide a more robust set of results. However, obtaining a large amount of ratings for a large amount of assets is very time consuming and randomising the process has the potential to result in an uneven distribution of ratings. The development of a method to address this aspect of large-scale listening tests is suggested as future work (10.3.2).

10.2 Contributions Of The Thesis

This thesis has presented several original contributions to emotional speech corpus construction, annotation and analysis, which made significant improvement to Mood Induction Procedures, corpus creation, speech annotation and listening tests.

1. An MIP was developed that was successful in eliciting natural underlying emotional speech from participants. This MIP used a game based imaginary survival scenario in which participants cooperated to complete a task, relying on false feedback regarding their progress to gauge how well they were doing. The MIP was a combination of the task-based false feedback MIP, the social interaction MIP and the gift giving MIP and was successful in avoiding demand effects and eliciting natural underlying emotional speech.

2.A HD audio environment was used for capturing underlying emotional speech. The devised MIP was recorded in a high quality audio environment using sound booths and professional level equipment. The audio recorded was of extremely high quality, noise free and presented a strong, clear, two-channel signal. These recordings were then archived and converted to a lower quality rate as needed for analysis and use for use in listening tests. Advancements in audio quality and recording equipment necessitates that audio be recorded at a high quality for archiving and to reduce data redundancy by future proofing the recorded speech against advancing technical standards.

3.A three-tiered annotation approach was proposed as means of comprehensively annotating emotional speech corpora. This took the form of metadata annotation, acoustic annotation and emotional annotation. A method of structuring the annotation and assets in a logical and coherent manner was implemented, which enabled the data to be accessed via a web-based emotional rating tool and an interactive visualisation front end. A three-tiered approach is necessary to comprehensively annotate emotional speech corpus assets. A backend database was shown to be necessary to structure the three levels of annotation in a coherent and logical manner. This addresses issues of data sparcity and ensures the reusability and usefulness of emotional speech corpora beyond their original remit. Furthermore, it enables researchers external to the creation of the corpus to understand and replicate data collection methods as well as carrying out a comparison of methodologies.

4.Large-scale web based listening tests were used to obtain a large amount of emotional dimensional ratings for corpus assets. This addressed the problems inherent in using small sample sizes and allowed for a more robust determination of the dimensional rating of each speech asset, which facilitated their subsequent acoustic analysis. The listening test methodology used provides a framework by which to carry out cross-cultural comparisons of emotional dimensional ratings and allows for long-term large-scale listening tests to be carried out.

5. A number of acoustic parameters were determined to be correlated with the activation dimension of emotion in relation to underlying emotional states. This substantiated some of the findings in the literature and demonstrated that there is a strong correlation between spectral energy distribution and the activation dimension of underlying emotional states. Furthermore it was determined that there was no correlation between the evaluation dimension of emotion in relation to underlying emotional states and certain acoustic parameters. These findings substantiate the findings in the literature and suggest a wider set of parameters needs to be considered in relation to each dimension.

This thesis has considered means by which a natural emotional speech corpus can be created, structured, annotated and analysed. In addition to the original contributions of this thesis, future work is required to fully develop the methods and techniques presented here.

10.3 Future Work

Work carried out in this thesis has produced significant improvement in the areas of emotional speech corpus construction, annotation, analysis, and emotional listening tests. A number of areas of investigation were undertaken and each is considered to determine how further development in these areas may proceed.

10.3.1 MIP Console Game Designed To Automatically Elicit Emotional Responses

An MIP was created to elicit emotional responses from participants. While the shipwreck MIP was successful in eliciting natural underlying emotional speech from participants, it relied on both inbuilt and external forms of manipulation. It would be advantageous to have a computer game with inbuilt emotional elicitation elements, with no need for external manipulation to be carried out. Computer games have been used in other research to elicit emotional responses and the case study gaming MIP was very positively received by participants. The creation of a purpose built computer game specifically designed to elicit emotional responses provides an ideal opportunity to create a MIP that is highly resistant to demand effects. With emotional elicitation

built into the game, no interaction or manipulation on the part of the researcher would be necessary. Coupled with a visually pleasing and immersive game that was easy to understand, an MIP using the game would be highly resistant to demand effects; the goals of the game would be the main focus of the participants and very little external manipulation would be evident. Furthermore the game could be created with a wide range of customisable options in order to tailor the game to certain demographics and participant abilities. An emotional elicitation game would also allow experiments to be replicated by other researchers possibly using different experimental groups, thus facilitating cross-cultural comparison to emotional stimuli.

10.3.2 Improvement Of The Large-Scale Listening Tests

A web-based tool was created to obtain a large number of ratings for the emotional speech assets. The method used for the shipwreck MIP attracted considerably more ratings due to the use of social networks. Future work will develop versions of the tool for use within a wide variety of social networks (such as a facebook plug-in). This should result in significantly more ratings and should address some of the issues raised regarding the limitations of the large scale listening tests. Dissemination of the listening tool throughout various social networks, with a robust and persistent backend database would allow a large-scale long term rating experiment to be carried out. With the pertinent aspects of the rating automated to a certain degree, a large amount of ratings could be obtained over the course of a year or more. Coupled with future work in relation to the automated annotation and analysis of emotional speech assets, an extremely strong set of rating results and analysis results could be obtained. Future work could also consider relating the emotional dimensional ratings of speech assets to categorical labels. This could be achieved by carrying out a three-stage set of listening tests. The first stage would obtain a large amount of dimensional ratings via the aforementioned long-term listening tests. The second stage of listening tests would obtain a large amount of user defined categorical labels for each asset. Analysis of these user-defined labels could then be used to create a list of the most commonly used labels for each asset. The final stage of listening tests would then use large scale listening tests to determine the most relevant categorical label for each asset based on the list compiled from the previous set of listening tests. Through the use of three-large scale listening tests a robust relationship between categorical labels and locations within a circumplex model could be determined.

10.3.3 Examining The Cultural Differences Among Listening Test Participants

Further work is needed to examine the cultural differences of users of the listening tool. The IP address of every user was collected and can be used to determine their country of origin. While the use of IP addresses in this manner is not without error, it does allow a rudimentary consideration of cultural differences. The creation of a listening tool plug-in for web based social networks allows for an increased amount of demographic data to be collected, thus providing a more accurate method of determining a users country of origin. This data can then be analysed to determine if users from different cultural backgrounds or geographic locations rate the assets in the corpus differently. This could then form the basis of an expanded set of rules for the use of web-based large-scale listening tests. It could also be used to further investigate whether acoustic expressions of emotion are universal in the same way as facial emotional displays.

10.3.4 Further Improving And Developing The Annotation Framework And Backend Database

Further work is needed to improve the annotation framework and persistent backend database underpinning it. While an adapted IMDI schema was successfully integrated into a backend database and corpus creation methodology, work is needed to improve and expand the metadata annotation. Future work will consider the possible integration of the emerging EmoML schema. Work is needed in the area of acoustic annotation. The use of the LinguaTag software demonstrated a method by which acoustic parameters could be stored and parsed into a backend database. The acoustic analysis achievable with LinguaTag was not comprehensive enough for this research and so the PRAAT analysis software was used. Ideally audio analysis of any sort should be stored in an XML document that can be parsed by a variety of database technologies. Future work will consider if it is possible to create software that can carry out a comprehensive analysis of an audio file and store it in a parsable file format or whether it is necessary to determine a fixed set of parameters for analysis. Further work will also consider the integration of an analysis procedure for the obtained emotional ratings. This would involve the determination of the central tendency and a confidence level for each asset and would drastically speed up the

analysis procedure by enabling a large part of the analysis to be semi-automated. This could then feed into an improved visualisation interface that could provide more detailed information and analysis results as part of the end user front-end.

10.3.5 Acoustic Analysis Of Underlying Emotion

Although a preliminary analysis of a small group of acoustic parameters was carried out in relation to underlying emotional states, the results indicate that a wider set of parameters needs to be considered, especially in relation to the evaluation dimension. While there is a consensus regarding certain core acoustic parameters in relation to a group of primary emotions, it is not certain that all these parameters are as important in relation to underlying emotional states. Improvements in this area will also entail a larger set of assets for future analysis, something that has already been addressed. The fact that the analysis carried out in chapter 9 resulted in some contradictory findings in relation to the literature in chapter 3 suggests that this needs to be investigated further. Ideally a set of different assets, subjected to large scale rating, would be analysed in the same manner to determine if the analysis results of chapter 9 can be replicated, while also considering a wider set of acoustic parameters. In particular, the further investigation of spectral energy in relation to underlying emotional states is needed to corroborate the relative findings in chapter 9.

10.3.6 Other Development

This research has investigated numerous aspects of emotional speech and the development of emotional speech corpora. Further work is needed in each area but work is also necessary regarding the use of subjective emotional term. A comprehensive definition of emotion that makes recourse to the physiological and neurological aspects as well as the social and evolutionary aspects is paramount. Any future definition must also consider the different levels of emotion, from primary and full-blown to underlying and less intense emotions. It is hoped that the development of a more rigorous definition of emotion will enable future research to proceed without having to first navigate the complex perspectives and definitions of emotion.

10.4 Overall Conclusions

10.4.1 Thesis Statement

This thesis was undertaken in relation to the following statement:

Determining the acoustic correlates of emotional speech presents numerous difficulties. While much work has been carried out in this area, there is no conclusive methodology or a definitive set of results upon which all commentators can agree. This is due to the complex subject, the wide ranging definitions of the term ‘emotion’ and its complex multidimensional nature. Examination of the acoustic correlates of emotional speech often relies on arguably substandard data and data collection methods, using actors or broadcast sources. This is further complicated through the use of emotional categories that have yet to be rigorously defined.

This statement was defended by answering six-research questions (10.1).

This research has considered the use of MIPs to elicit natural underlying emotional speech and the annotation and structuring of speech corpus assets. Significant improvements were produced during the development of an MIP to elicit natural underlying emotional speech, the annotation and structure of speech corpus assets, the use of a web-based listening test to obtain a large amount of ratings and the acoustic analysis of underlying emotional speech. Although further work is required in all of these areas (10.3), the contributions of this thesis have gone some way towards a comprehensive method of obtaining, annotating, storing and analysing natural underlying emotional speech assets.

11. Bibliography

(IEC), I. E. C. (1999). INTERNATIONAL STANDARD 60908: Audio recording – Compact disc digital audio system.

Abelin, A. and J. Allwood (2000). Cross Linguistic Interpretation of Emotional Prosody. ISCA workshop on Speech and Emotion Northern Ireland.

Abelson, R. P., AND Sermat, Vello. (1962). "MULTIDIMENSIONAL SCALING OF FACIAL EXPRESSIONS." Journal of Experimental Psychology **63**(6): 546-554.

Adobe. (2009). "Adobe Flex Home Page." from http://www.adobe.com/products/flex/.

Alter, K., E. Rank, et al. (2000). Accentuation and emotions - two different systems? ISCA workshop on speech and emotion, Northern Ireland.

Amir, N., S. Ron, et al. (2000). Analysis of an emotional speech corpus in Hebrew based on objective criteria. ISCA ITRW on Speech and Emotion, Belfast.

Anttonen, J., V. Surakka, et al. (2009). "Ballistocardiographic responses to dynamic facial displays of emotion while sitting on the eMFi chair." Journal of Media Psychology: Theories, Methods, and Applications **21**(2): 69-84.

Apple-Computers. (2009). "Apple Mac Home Page." from http://www.apple.com/mac.

Arnold, M. B. (1960). Emotion and personality. New York, Columbia University Press.

Arnott, I. R. M. J. L. (1995). "Implementation and testing of a system for producing emotion-by-rule in synthetic speech." Speech Communication **16**(4): 369-390.

Association, B.-R. D. (2005). White paper: Blu-ray Disc Format**:** 35.

Averill, J. R. (1980). A constructivist view of emotion. Emotion: Theory, research and experience. R. Plutchik and H. Kellerman. New York, Academic Press. **1:** 305-339.

Avid-Technology. (2007). "Avid Pro-Tools website." 2007, from http://www.avid.com/US/resources/digi-orientation.

Ax, A. (1953). "The physiological differentiation between fear and anger in humans." Psychosomatic Medicine **55** (5): 433–442.

Baggia, P., F. Burkhardt, et al. (2009). "Emotion Markup Language (EmotionML) 1.0." 2009, from http://www.w3.org/TR/emotionml/.

Baken, R. J. and R. F. Orlikoff (2000). Clinical measurement of speech and voice, Cengage Learning.

Banos, R. M. B., C. ; Liaño. V. ; Rey, B. ; Guerrero, B., Alcaniz, M. (2003). Virtual Reality as Mood Induction Procedure. PRESENCE 2003, 6th Annual International Workshop on Presence.

Banse, R. and K. R. Scherer (1996). "Acoustic profiles in vocal emotion expression." Journal of Personality and Social Psychology **70**(3): 614-636.

Banziger, T., V. Tran, et al. (2005). The Geneva Emotion Wheel: A tool for the verbal report of emotional reactions. ISRE 2005, Conference of the International Society for Research on Emotions, Bari, Italy.

Barsade, S. G. (2002). "The Ripple Effect: Emotional Contagion and its Influence on Group Behavior." Administritive Science Quarterly **47**(4): 644-675.

Baumgardiner, A. H., Arkin, R.M. (1988). "Affective state mediates causal attributons for success and failure." Motivation and Emotion **12**: 99-111.

Beedie, C. J., Terry, P.C., Lane, A. M. (2005). "Distinctions between emotion and mood." Cognition and Emotion **19**(6): 847-878.

Bellman, B. L., & Bennetta Jules-Rosette. (1977). A Paradigm for looking. Norwood, Ablex Publishing.

Biersack, S., & Kempe, V. (2005). Tracing vocal emotion expression through the speech chain: Do listeners perceive what speakers feel? ISCA Workshop on Plasticity in Speech Perception, London UK.

Blu-Ray-Association (2005). White paper: Blu-ray Disc Format. **35**.

Boersma, P. and D. Weenink (2006). Praat: doing phonetics by computer.

Boslaugh, S., & Watters, Paul. A. (2008). Statistics in a nutshell. Sebastopol, O'Reilly Media, Inc.

Bradley, M., Greenwald, M.K., Hamm, A.O (1993). Afffective picture processing The Structure of Emotion. N. Birbaumer, Ohman, A. Toronto**:** 48-65.

Brave, S. and C. Nass (2008). Emotion in Human-Computer Interaction The Human-Computer Interaction Handbook: Fundamentals, Evolving Techologies and Emerging Applications. A. Sears and J. A. Jacko. New York, Taylor Francis Group.

Breazeal, C. L. (2004). Designing Sociable Robots, The MIT Press

Brockmann, M., C. Storch, et al. (2008). "Voice Loudness and Gender Effects on Jitter and Shimmer in Healthy Adults." Journal of Speech, Lnaguage, and Hearing Research **51**: 1152-1160.

Bryman, A. and D. Cramer (2004). Quantitative Data Analysis with SPSS 12 and 13: A Guide for Social Scientists, Routledge.

Buck, R. (1999). "Biological affects: A typology." Psychological Review(106): 301-336.

Cahn, J. (1990). Generating Expression in Synthesized Speech MIT Media Laboratory. Boston, MIT. **Masters**.

Campbell, N. (2000). Databases of emotional speech. ISCA Workshop on Speech and Emotion, Northern Ireland.

Cappella, J. N. (1993). "The facial feedback hypothesis in human interaction." Journal of Language and Social Psychology(12): 13-29.

Christie, I. C. (2002). Multivariate Discrimination of Emotion-Specific Autonomic Nervous System Activity. . Science, Virginia Polytechnic Institute and State University. **MSc**.

Chung, S. (1999). Vocal expression and perception of emotion in Korean. 14th International Conference of Phonetic Sciences, San Fransisco, USA.

Cook, P. R. (2001). Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics, The MIT Press %@ 0262531909.

Core, D. (2009). "Dublin Core Metadata Initiative." from http://dublincore.org/specifications/.

Cornelius, R. R. (1996). The Science of Emotion. Research and Tradition in the Psychology of Emotion. Upper Saddle River, NJ, Prentice-Hall.

Cornelius, R. R. (2000). "Theoretical approaches to emotion." Speech Emotion **1**: 3-10.

Cornelius, R. R. (2000). Theoretical Approaches to Emotion. ISCA Workshop on Speech and Emotion, Belfast, northern Ireland.

Cowie, R. and R. R. Cornelius (2003). "Describing the emotional states that are expressed in speech." Speech Communication Special Issue on Speech and Emotion **40**(1-2): 5-32.

Cowie, R., E. Douglas-Cowie, et al. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. ISCA Workshop on Speech and Emotion, Northern Ireland.

Cowie, R., E. Douglas-Cowie, et al. (2001). "Emotion recognition in human-computer interaction." IEEE Signal Processing Magazine **18**(1): 32-80.

Cullen, C., B. Vaughan, et al. (2008a). LinguaTag: an emotional speech analysis application. Accepted paper at: The 12th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2008. Orlando, Florida, USA.

Cullen, C., B. Vaughan, et al. (2008c). Emotional Speech Corpus Construction, Annotation and Distribution. The 6th edition of the Language Resources and Evaluation Conference. Marrakech (Morocco).

Cullen, C., B. Vaughan, et al. (2008). Emotional Speech Corpora for Analysis and Media Production. 3rd International Conference on Semantic and Digital Media Technologies, SAMT. Koblenz, Germany.**:** 2.

Cullen, C., B. Vaughan, et al. (2008b). A vowel-stress emotional speech analysis method. CITSA. Genoa, Italy.

Cullen, C., Vaughan, B. ,Kousidis, S., Wang, Yi ., McDonnell, C. and Campbell, D. (2006). Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction International Conference on Multidisciplinary Information Sciences and Technologies Extremadura, Merida.

Cullen, C., Vaughan, B., Mc Auley, J., & Mc Carthy, E. (2009). "CorpVis: An Online Emotional Speech Corpora Visualisation Interface." Semantic Multimedia **5887**: 169-172.

Cunningham, M. R., Schaffer, D. R., Barbee, A.P., Wolff, P.L,. Kelley, D.J (1990). "Seperate processes in the relation of elation and depression to helping : Social versus personal concerns " Journal of Experimental Social Psychology **26**: 13-33.

Cutler, A., D. Dahan, et al. (1997). "Prosody in the Comprehension of Spoken Language: A Literature Review." Language and Speech(40(2)): 141-201.

Darwin, C. R. (1872). The Expression of the Emotions in Man and Animals. London., Albermarle.

Davis, J. I., Senghas, A., and Ochsner, K. N. (2009). "How does facial feedback modulate emotional experience? ." Journal of Research in Personality **43**: 822-829.

De Jong, N. H. and T. Wempe (2009). "Praat script to detect syllable nuclei and measure speech rate automatically " Behavior Research Methods **41**(2): 385-390.

Dejonckere, P., M. Remacle, et al. (1996). "Differentiated perceptual evaluation of pathological voice quality: Reliability and correlations with acoustic measurements. ." Revue de laryngologie, d'otologie et de rhinologie, **117**(3): 219–224.

Detenber, B. H., Simons, R.F., Bennett., G.G (1998). "Roll 'em!: the effects of picture motion on emotional responses." Journal of Broadcasting and Elerctronic Media **21**: 112-126.

Dietz, R. B. and A. Lang (1999). Æffective Agents: Effects of Agent Affect on Arousal, Attention, Liking & Learning. 3rd International Cognitive Technology Conference, San Francisco, CA, USA.

Digidesign. (2009). "Digidesign's Pro-Tools HD specifications page.", from http://www.digidesign.com/index.cfm?langid=1&itemid=4892.

Douglas-Cowie, E., N. Campbell, et al. (2003). "Emotional speech: towards a new generation of databases." Speech Communication Special Issue Speech and Emotion **40**(1-2): 33–60.

Douglas-Cowie, E., R. Cowie, et al. (2000). A new emotion database: considerations, sources and scope. ISCA Workshop on Speech and Emotion, Northern Ireland.

Dowling, W. J. and J. L. Harwood (1985). Music Cognition, Academic Press.

DSM-IV, T. F. o. (1994). Diagnostic and Statistical Manual of Mental Disorders (4th Edition) (DSM-IV). Washington, DC, American Psychiatric Association.

Dupoux, E., N. Sebastia´n-Galle´s, et al. (2008). "Persistent stress 'deafness': The case of French learners of Spanish." Cognition **106**(2): 25.

Dutoit, T. (1997). An Introduction to Text-to-Speech Synthesis. Dordrecht, Kluwer Academic Publishers.

Edgington, M. (1997). Investigating the limitations fo concatenative synthesis. Eurospeech 97, Rhodes, Greece.

Ekman, P. (1993). "Facial expression and emotion." American Psychologist **48**(4): 384-392.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A.,Diacoyanni-Tarlatzis, I., Heider, K., Krause, R.,LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E.,Scherer, K. R., Tomita, M., & Tzavaras, A. (1987). "Universals and cultural differences in the

judgments of facial expressions of emotion." Journal of Personality and Social Psychology **53**(4): 712-717.

Ekman, P. and E. L. Rosenberg (2005). What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Oxford University Press, USA %@ 0195179641 %7 2.

Enberg, I. S., Hansen, A.V., Anderson, O., Dalsgaard, P. (1997). Design, recording and verification of a Danish Emotional Speech Database. Eurospeech '97, Rhodes, Greece.

Erickson, F., and Schultz, J. (1982). The Counsellor as Gatekeeper: Social Interaction in Interviews. Language, Thought and Culture: Advances in the Study of Cognition. E. Hammel. New York, Academic Press.

F. Pellegrino., R. A.-O. (1999). An unsupervised approach to language identification. IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE Computer Society. **2**.

Facebook. (2010). "Facebook User Statistics." from http://www.facebook.com/press/info.php?statistics.

Farmer, A., D. Lam, et al. (2006). "A pilot study of positive mood induction in euthymic bipolar subjects compared with healthy controls." Psychological Medicine **36**: 1213-1218.

Fernandez, R. and R. Picard (2000). Modelling drivers' speech under stress. ISCA Workshop on Speech and Emotion, Northern Ireland.

Fitzgerald, M. J. T., Folan-Curran, Jean. (2002). Clinical Neuroanatomy and Related Neuroscience, W. B. Saunders.

Fontaine, J. R., K. R. Scherer, et al. (2007). "The world of emotion is not two-dimensional." Psychological Science **18**(2): 1050,Äì1057.

Forgras, J. P. (1990). "Affective influences on individual and group judgements ." European Journal of Social Psychology **20**: 441-453.

Fredrickson, B. L., Mancuso, R.A., Branigan, C., Tugade, M.M. (2000). "The undoing effect of positive emotions." Motivation and Emotion **24**(4): 237–257.

Frijda, N. H. (1986). The Emotions. Cambridge, UK, Cambridge University Press.

Frijda, N. H. (1994). Varieties of affect: Emotions and episodes, moods, and sentiments. The nature of emotion. P. Ekman and R. J. Davidson. New York, Oxford University Press.**:** 59-67.

Gazzaniga, M. S., Richard., I., B.,l George, R. M. (2009). Cognitive Neuroscience: The Biology of the Mind. New York, W.W Norton & Company, Inc.

Geer, B. a. H. S. B. (1957). "Participant Observation and Interviewing: A Comparison." Human Organization **16**(3): 28-32.

Gerrards-Hesse, A., K. Spies, et al. (1994). "Experimental inductions of emotional states and their effectiveness: A review." British Journal of Psychology **85**(1): 55-78.

Gobl, C., E. Bennett, et al. (2002). Expressive Synthesis: How Crucial is Voice Quality? IEEE Workshop on Speech Synthesis, Santa Monica, CA (USA).

Gobl, C. and A. Ní Chasaide (2000). Testing Affective Correlates of Voice Quality Through Analysis and Resynthesis. ITRW on SPeech and Emotion. Newcastle, Northen Ireland, UK**:** 6.

Gobl, C. and A. Ní Chasaide (2003). "The Role of voice quality in communicating emotion, mood and attitude." Speech Communication **40**: 189-212.

Goritz, A. S. (2007). "The Induction of Mood via the WWW." Motivation and Emotion **31**: 35-47.

Gottdiener, M. (1979). "Field Research and Video Tape." Sociological Inquiry **4**(49): 59-66.

Greenwald, M. K., E. W. Cook, et al. (1989). "Affective judgment and psychophysiological response: Dimensional covariation in the evalua- tion of pictorial stimuli,." Journal of Psychophysiology **3**: 51–64.

Gross, J. J. and R. W. Levenson (1995). "Emotion elicitation using films." Cognition and Emotion(9): 87-108.

Group, T. P. (2009). "Home page of The PHP Group." from http://php.net/index.php.

Hacker, S. (2000). MP3: The Definitive Guide, O ' Reilly.

Hanjalic, A., Xu, et al. (2005). "Affective Video Content Representation and Modeling." IEEE Transactions on Multimedia **7**(1): 143-154.

Hansson, D. H. (2009). "Ruby-on-Rails home page." from http://rubyonrails.org/.

Harrington, J. and S. Cassidy (1999). Techniques in Speech Acoustics, Springer

Henkel, J. M. and V. B. Hinsz (2004). "Success and Failure in Goal Attainment as a Mood Induction Procedure." Social Behavior and Personality **32**(8): 715-722.

Hill, T. and P. Lewicki (2005). Statistics: Methods and Applications, StatSoft, Inc.

Holdener, A. (2008). Ajax: The Definitive Guide, O' Reilly Media.

Honorof, D. N. and D. H. Whalen (2005). "Perception of pitch location within a speaker's F0 range." The Journal of the Acoustical Society of America **117**(4): 2193-2200.

Howard, D. and J. Angus (1999). Acoustic and Psychoacoustics, Focal Press.

Huber, D. M., Runstein, Robert. E. (2005). Modern Recording Techniques, Focal Press.

IBM. (2010). "Home page for PASW/SPSS software." from http://www.spss.com/statistics/.

IEC. (2009). "IEC Homepage." from http://www.iec.ch/zone/si/si_bytes.htm.

IEC, I. E. C. (1999). INTERNATIONAL STANDARD 60908: Audio recording – Compact disc digital audio system.

Iida, A., N. Campbell, et al. (1998). "Design and Evaluation of Synthesised Speech with Emotion." Journal of Information Processing Society of Japan **40**(2): 479-486.

Iida, A., Campbell, N., Higuchi, F., Yasumura, M. (2003). "A corpus-based speech synthesis system with emotion." Speech Communication **40**: 26.

IMDI. (2007). "IMDI Home Page." from http://www.mpi.nl/IMDI/.

IMDI. (2007). "IMDI Metadata Domain." 2009, from http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI240467%23.

Iriondo, I., Guaus, R., Rodriguez, A. (2000). VALIDATION OF AN ACOUSTICAL MODELLING OF EMOTIONAL EXPRESSION IN SPANISH USING SPEECH SYNTHESIS TECHNIQUES. ITRW on Speech and Emotion. Newcastle, Northern Ireland, UK, ISCA Archive.

Jack, R. E., C. Blais, et al. (2009). Cultural Confusions Show that Facial Expressions Are Not Universal. Current Biology. **19:** 1543-1548.

Jain, R. (2004). Quality of Experience. IEEE Multimedia.

James D. Laird., J. J. W., Mark Halal., and Martha Szegda (1982). "Remembering What You Feel: Effects of Emotion on Memory." Journal of Personality and Social Psychology" **45**(4): 646-657.

James, W. (1884). "What is an emotion?" Mind **9**: 188-205.

Jansz, J. (2005). "The Emotional Appeal of Violent Video Games for Adolescent Males." Communication Theory **15**(3): 219-241.

Johns-Lewis, C. (1986). Prosodic differentiation of discourse modes. Intonation in Discourse. C. Johns-Lewis. San-Diego, College Hill Press**:** 199-220.

Johnstone, T. (1996). Emotional Speech Elicited Using Computer Games. 4th International Conference on Speech and Language Processing, Philadelphia, PA, USA.

Johnstone, T., C. M. Reekum, et al. (2005). "Affective Speech Elicited With a Computer Game." Emotion, 2005 **5**(4): 513-518.

Johnstone, T., C. M. v. Reekum, et al. (2005). "Affective speech elicited with a computer game." Emotion(5): 513-518.

Johnstone, T. and K. R. Scherer (1999). The Effects of Emotions on Voice Quality. XIV Int. Congress of Phonetic Sciences, San Francisco.

Juslin, P. N. and P. Laukka (2003). "Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?" Psychological Bulletin **129**(5): 770-814.

Juul, J. (2003). The Game, the Player, the World: Looking for a Heart of Gameness. Level Up: Digital Games Research Conference Proceedings. M. C. a. J. Raessens, Utrecht University, 2003.**:** 30-45.

Kaiser, S. and T. Wehrle (2000). Ausdruckspsychologische Methoden. Weinheim, Beltz, Psychologie Verlags Union.

Kaiser, S., T. Wehrle, et al. (1994). Multi-modal emotion measurement in an interactive computer-game: A pilot-study. . The VIIIth Conference of the International Society for Research on Emotions (

Kaiser, S. W., T. (1996). Situated emotional problem solving in interactive computer games. VIXth Conference of the International Society for Research on Emotions, Storrs, Connecticut ISRE Publication.

Kandel, E., Schwartz, J., Jessell, T. (2000). Principles of Neural Science, McGraw-Hill Medical.

Katz, B. (2002). Mastering Audio: The Art and the Science. Burlington, MA, Focal Press.

Kehrein, R. (2002). The prosody of authentic emotions. Speech Prosody, Aix-en-Provence, France.

Kienast, M., Sendlmeier, W.F., (2000). Acoustical analysis of spectral and temporal changes in emotional speech. ISCA ITRW on Speech and Emotion, Newcastle, Belfast, Textflow.

Kirchsteiger, G., L. Rigotti, et al. (2006). "Your morals might be your moods." Journal of Economic Behaviour & Organization **59**: 155-172.

Kochanski, G. (2006). Prosody Beyond Fundamental Frequency. Methods in Empirical Prosody Research S. Sudhoff, D. Lenertova, R. Meyeret al. Berlin, Walter de Gruyter.

Kooijman. V, H., P ,Cutler. A (2005). "Electrophysiological evidence for prelinguistic infants' word recognition in continuous speech." Cognitive Brain Research **24** 109–116.

Kotropoulos, D. V. a. C. (2003). A State of the Art Review on Emotional Speech Databases. 1st Richmedia Conference. Laussane**:** 109-119.

Kousidis, S., Dorran, D., McDonnell, C., Coyle, E. (2009). Time Series Analysis of Acoustic Feature Convergence in Human Dialogues. 13th International Conference on Speech and Computer (SPECOM). St. Petersburg, Russia.

Laukka, P., P. N. Juslin, et al. (2005). "A dimensional approach to the vocal expression of emotion." Cognition and Emotion **19**(5): 633-653.

Laver, J. (1994). Principles of Phonetics. Cambridge, Cambridge University Press.

Lazarus, R. (1991). Emotion and Adaptation, Oxford University Press.

Lazarus, R. S. (1999). The cognition-emotion debate: a bit of history. Handbook of Cognition & Emotion. T. Dalgleish and M. J. Power. New York, John Wiley**:** 3-19.

Lehiste, I. (1970). Suprasegmentals. Cambridge, MA, MIT Press.

Lerdorf, R., MacIntyre, P., & Tatroe, K. (2006). Programming PHP, 2nd Edition. Sebastopol, O' Reilly Media.

Malatesta, L., J. Murray, et al. (2009). Emotion Modelling and Facial Affect Recognition in Human-Computer and Human-Robot Interaction. State of the Art in Face Recognition. I. C. M. Dr. Mario.

Mansoorizadeh, M., Charkarie, M. Nasrollah. (2007). Speech emotion recognition: Comparison of speech segmentation approaches. Third information and Knowledge Technology(IKT07). Ferdowsi University of Mashhad, Mashhad, Iran**:** 5.

Manucia, G. K. B., D. J. & Cialdini, R. B. (1984). "Mood influences on helping: Direct effects or side effects?" Journal of Personality and Social Psychology **46**: 357-364.

Matsumoto, D., B. Willingham, et al. (2009). "Sequential Dynamics of Culturally Moderated Facial Expressions of Emotion." Psychological Science **20**(10): 1269-1275.

Mauss, I.B., et al. (2009). "Measures of emotion: A review." Cognition and Emotion **23**(2): 209-237.

Mehrabian, A. a. R., J.A. (1974). An Approach to Environmental Psychology Cambridge, MIT Press.

Microworkers.com. (2009). "Home page of Microworkers.com." 2010, from http://www.microworkers.com/.

Millenson, J. R. (1967). Principles of Behavioral Analysis, Prentice Hall.

Mixdorff, H. (2002). Speech Technology, ToBI, and Making Sense of Prosody. . Aix-en-Provence, France.

Mozziconacci, S. j. L., Hermes, D.J. (1997). A Study of Intonation Patterns in Speech Expressing Emotion or Attitude: Production and Perception. IPO Annual Progress Report. Eindhoven. **32**.

MPEG. (2004). "Mpeg-7 Overview." from http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

Mueller, J. S. and J. R. Curhan (2006). "Emotional intelligence and counterpart mood induction in a negotiation " International Journal of Conflict Management **17**(2): 110-128.

Murray, I. R. A., John L. (1993). "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion." The Journal of the Acoustical Society of America **93**(2): 1097-1108.

Netter, F. H. (1997). Atlas of Human Anatomy Rittenhouse.

Neumann. (2009). "Neumann Microphones webpage." from http://www.neumann.com/?lang=en&id=about_us_overview.

Niimi, Y., Kasamatu, M.L., Nishimoto, T., & Araki, M. (2001). Synthesis of emotional speech using prosodically balanced VCV segments. ISCA Tutorial & Workshop on Research Synthesis Scotland.

Noble, J., & Anderson, T. (2008). Flex 3 Cookbook, O'Reilly Media.

Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. . Proceedings of the 15th International Congress of Phonetic Science., Barcelona, Spain. .

Nummenmaa, L. and P. Niemi (2004). "Inducing Affective States With Success-Failure Manipulations: A Meta-Analysis." Emotion **4**(2): 207-214.

Nyquist, H. (2002). "Certain Topics in Telegraph Transmission Theory." PROCEEDINGS OF THE IEEE, **90**(2): 280-305.

O'Reilly, M. and A. Ní Chasaide (2007). Analysis of Intonation Contours in Portrayed Emotions Using the Fujisaki model. The 2nd International Conference on Affective Computing and Intelligent Interaction. R. Cowie, de Rosis, F. Lisbon, Portugal**:** 102-109.

Oehme, A., Hebron, A., Kupschick, S. & Zentsch, E. (2007). Physiological Correlates of Emotions. AISB Annual Convention, Newcastle upon Tyne, UK.

Oertel, K., Fischer, G., Diener, H. (2004). Physiological Response to Games and Non-games: A Contrastive Study. Entertainment Computing – ICEC 2004. Berlin, Springer Berlin / Heidelberg. **3166:** 402-405.

OLAC. (2008). "Open Language Archives Community home page." from http://www.language-archives.org/.

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

Osgood, C. E. (1960). "Cognitive Dynamics in the Conduct of Human Affairs." The Public Opinion Quarterly **24**(2): 341-365.

Osgood, C. E., G. J. Suci, et al. (1957). The measurement of meaning. Urbana, USA, University of Illinois Press.

Oudeyer, P. (2003). "The production and recognition of emoitons in speech: features and algorithims." International Journal of Human-Computer Studies **59**: 157-183.

Owren, M. J. and J. Bachorowski (2007). Measuing Emotion-Related Vocal Acoustics. Handbook of Emotion Elicitation and Assessment, Oxford University Press, USA

Ozdas, A., Shiavi, R.G., Silverman, S.E., Silverman, M.K., Wilkes, D.M. (2004). "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.

Papez, J. W. (1937). "A Proposed Mechanisim of Emoiton." Archives of Neurological Psychiatry **38**: 725-743.

Parkinson, B., Totterdell, P., Briner. R. B., & Renyolds, S. (1996). Changing Moods: The Psychology of Mood and Mood Regulation. Harlow, UK., Addison Wesley Longman.

Patel, A. D. (2006). "An Empirical Method for Comparing Pitch Patterns in Spoken and Musical Melodies: A Comment on J.G.S. Pearl's "Eavesdropping with a

Master: Leos Janá ek and the Music of Speech." Empirical Musicology review **1**(3): 166-169.

Peat, J. K. and B. Barton (2005). Medical statistics, Wiley-Blackwell

Pereira, C. (2000). Dimensions of emotional meaning in speech. ISCA ITRW on Speech and Emotion, Newcastle, Belfast, Textflow.

Peter, C. and A. Herbon (2006). "Emotion representation and physiology assignments in digital systems." Interacting with Computers **18**(2): 139-170.

Picard, R. W., E. Vyzas, et al. (2001). "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State." IEEE Trans Pattern Analysis & Machine Intelligence(23): 1175-1191.

Plutchik, R. (1980). Emotion: A psychoevolutionary synthesis. New York, Harper and Row.

Plutchik, R. (1994). The Psychology and Biology of Emotion. New York, HarperCollins College Publishers.

Plutchik, R. (2001). "The Nature of Emotions." American Scientist **89**(4): 344-350.

Pohlmann, K. (2000). Principles of Digital Audio, McGraw-Hill.

Polzin, T. S., Waibel, A. (2000). Emotion-sensitive human-computer interfaces. ISCA ITRW on Speech and Emotion Newcastle, Belfast, UK.

PRAAT-Users-Group. (2010). "PRAAT users group." from http://uk.groups.yahoo.com/group/praat-users/.

Pugmire, D. (1994). "Real Emotion." Philosophy and Phenomenological research **54**(1): 105-122.

Puri, B. K., Tyrer, P. J. (1998). Sciences Basic to Psychiatry, Churchill Livingstone.

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. Annual Review of Language Acquisition 2 John Benjamins Publishing Company. **2**.

Ramus, F., M. Nespor, et al. (1999). "Correlates of linguistic rhythm in the speech signal." Cognition **73**(3): 265-292.

Rehm. M., André. E., et al. (2007). The CUBE-G approach – Coaching culture-specific nonverbal behavior by virtual agents. . In Proceedings of the 38th Conference of the International Simulation and Gaming Association (ISAGA). Nijmegen - The Netherlands.

Rizzo, P. (2001). Why should agents be emotional for entertaining users? A critical analysis. Affective interactions: towards a new generation of computer interfaces New York, Springer-Verlag**:** 166-181.

Roach, P. (2000). Techniques for the phonetic description of emotional speech. ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland. .

Roach, P., R. Stibbard, et al. (1998). "Transcription of Prosodic and Paralinguistic Features of Emotional Speech." Journal of the International Phonetic Association(28): 83-94.

Roads, C. (1996). The Computer Music Tutorial, MIT Press.

Robinson, M. D. and G. L. Clore (2002). "Episodic and Semantic Knowledge in Emoitonal Self-Report: Evidence for Two Judgement Processes." Journal of Personality and Social Psychology **83**(1): 198-215.

Roederer, J. G. (2008). The Physics and Psychophysics of Music: An Introduction, Springer 4th ed.

Rottenberg, J., R. R. Ray, et al. (2007). Emotion elicitation using films. The handbook of emotion elicitation and assessment. J. A. Coan and J. J. B. Allen. New York, Oxford University Press.

Ruby, J. (1980). "Exposing yourself: Reflexivity, anthropology, and film." Semiotica: The Journal of the International Association for Semiotic Studies **30**(1-2): 153-179.

Ruffle, B. J. (1999). "Gift giving with emotions." Journal of Economic Behaviour & Organization **39**: 399-420.

Rumsey, F. and T. McCormick (2002). Sound and Recording: An Introduction, Third Edition (Music Technology) Burlington, MA, Focal Press.

Russ, M. (2008). Sound Synthesis and Sampling, Focal Press.

Russell, J. A. (1980). "A circumplex model of affect." Journal of Personality and Social Psychology(39): 1161-1178.

Russell, J. A. (1993). "Forced-choice response format in the study of facial expression. ." Motivation and Emotion **17**: 41 -51.

Russell, J. A. (1994). "Is there universal recognition of emotion from facial expressions? A review of cross-cultural studies." Psychological Bulletin(115): 102-141.

Russell, J. A. (1997). How shall an emotion be called?, American Psychological Association (APA)

Russell, J. A., J. Bachorowski, et al. (2003). "Facial and Vocal Expressions of Emotion." Annual Review of Psychology **54**: 329-349.

Sabini, J. and M. Silver (2005). "Why Emotion Names and Experiences Don't Neatly Pair." Psychological Inquiry **16**(1): 1-10.

Sánchez, A. j., N. P. Hernández, et al. (2006). Conveying mood and emotion in instant messaging by using a two-dimensional model for affective states. VII Brazillian Symposium on Human Factors in Computing Systems, Natal, R, Brazil.

Schachter, S., Singer, J. (1962). " Cognitive, social, and physiological determinants of emotional state." Psychological Review **69**(5): 379-399.

Scherer, K. (1972). Acoustic Concomitants of Emotional Dimensions: Judging Affect from Synthesized Tone Sequences. Eastern Psychological Association Meeting. Boston, Massachusetts.

Scherer, K. (1981a). Speech and Emotional States Speech evaluation in psychiatry. New York, Grune & Stratton**:** 31.

Scherer, K. R. (1979). Non-linguistic vocal indicators of emotion and psychopathology. New York, Plenum Press.

Scherer, K. R. (1981b). Speech and emotional states. New York, Grune & Stratton.

Scherer, K. R. (1984a). On the nature and function of emotion: A component process

approach. Approaches to emotion. K. R. Scherer and P. Ekman. Hillsdale, NJ, Erlbaum**:** 293-317.

Scherer, K. R. (1984b). "Emotion as a multicomponent process: A model and some crosscultural

data." Review of Personality and Social Psychology **5**: 37-63.

Scherer, K. R. (1986b). "Vocal affect expression: A review and a model for future research." Psychological Bulletin **99**: 143– 165.

Scherer, K. R. (2000e). Psychological models of emotion. Oxford/New York, Oxford University Press.

Scherer, K. R. (2001a). Appraisal considered as a process of multi-level sequential checking. New York and Oxford, Oxford University Press.

Scherer, K. R., R. Banse, et al. (2001b). "Emotion inferences from vocal expression correlate across languages and cultures." Journal of Cross-Cultural Psychology **32**(1): 76-92.

Scherer, K. R. and G. Ceschi (1997). "Lost luggage: A field study of emotion-antecedent appraisal." Motivation and Emotion **21**: 211-235.

Scherer, K. R. and G. Ceschi (2000c). "Criteria for emotion recognition from verbal and nonverbal expression: Studying baggage loss in the airport." Personality and Social Psychology Bulletin **26**(3): 327-339.

Scherer, K. R., E. Dan, et al. (2006). "What determines a feeling's position in three-dimensional affect space? A case for appraisal." Cognition and Emotion **20**(1): 92-113.

Scherer, K. R., T. Johnstone, et al. (2003). Vocal Expression of Emotion. New York and Oxford, Oxford University Press.

Schlosberg, H. (1941). "A scale for the judgement of facial expressions." Journal of Experimental Psychology **29**: 497-510.

Schlosberg, H. (1952). "The description of facial expressions in terms of two dimensions." Journal of Experimental Psychology **44**: 229-237.

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

Schneider, K. and I. Josephs (1991). "The expressive and communicative functions of preschool childrens smiles in an achievment-situation." Journal of Nonverbal behaviour **15**(3): 185-198.

Scholsberg, H. (1941). "A Scale For The Judgment of Facial Expressions." Journal of Experimental Psychology **29**: 497-510.

Schröder, M. (2004a). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. Workshop on Affective Dialogue Systems, Kloster Irsee, Germany.

Schröder, M. (2004b). Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis. Faculty of Philosophy, Universitat des Saarlandes**:** 288.

Schröder, M., R. Cowie, et al. (2001). Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. 7th European Conference on Speech Communication and Technology-2nd INTERSPEECH Event, Aalborg, Denmark.

Schröder, M., H. Pirker, et al. (2006). First Suggestions for an Emotion Annotation and Representation Language.

Shami, M., Kamel, Mohammad (2006). Segment-based approach to the recognition of emotions in speech. 2005 IEEE International Conference on Multimedia and Expo. Amsterdam, Netherlands**:** 2.

Sharma, A. K. (2005). Text Book of Elementary Statistics. Delhi, India, Discovery Publishing House.

Shaver, P., J. Schwartz, et al. (1987). "Emotion Knowledge: Further exploration of a prototype approach." Journal of Personality and Social Psychology **52**(1): 1061-1086.

Shaver, P. R., S. Wu, et al. (1992). "Cross-cultural similarities and differences in emotion and its representation: A prototype approach."

Shiota, M. N., Kletner, D. (2005). "What Do Emotion Words Represent?" Psychological Inquiry **16**(1): 32-37.

Sinnatamby, C. S. (1999). Last's Anatomy: Regional and Applied, Churchill Livingstone.

Sloman, A. (1998). Damasio, Descartes, Alarms and Metamanagement IEEE International Conference on Systems, Man, and Cybernetics., San Diego.

Sloman, A. a. C., R. and Scheutz, M. (2005). The Architectural Basis of Affective States and Processes Who Needs Emotions?: The Brain Meets the Robot. . New York, USA., Oxford University Press**:** 203-244.

SMIL, W. W. W. C. (2008). "Synchronized Multimedia." from http://www.w3.org/AudioVideo/.

Software, E. (2008). "Eiffel Software Home Page." Retrieved February, 2008.

Sonic-Solutions (2000). Understanding DVD Audio: A Sonic White Paper. Novato, California., Sonic Solutions. **17**.

Spackman, M. P., B. L. Brown, et al. (2009). "Do emotions have distinct vocal profiles? A study of idiographic patterns of expression " Cognition and Emotion **23**(8): 1565-1588.

Spanias, A., T. Painter, et al. (2007). Audio Signal Processing and Coding, Wiley-Interscience.

Spiesman, E. V., Lazarus, R.S., Mordkoff, A., & Davison, L. (1964). "Experimental reduction of stress based on ego-defense theory." Journal of Abnormal and Social Psychology **68**: 367-380.

Steidl, S. (2009). Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. Engineering. Nürnberg, University of Erlangen-Nürnberg. **Ph.D:** 260.

Stemmler, G. (1992). "The vagueness of specificity: Models of peripheral physiological emotion specificity in emotion theories and their experimental discriminability." Journal of Psychophysiology **6**: 17-28.

Stemmler, G. H., M., Pauls, C.A., Scherer, T. (2001). "Constraints for emotion specificity in fear and anger: The context counts." Psychophysiology(38): 275-291.

Strack, F., Schwarz, N., Gschneidinger, E. (1985). "Happiness and reminiscing: The role of time perspective, effect, and mode of thinking." Journal of Personality and Social Psychology **49**: 1460-1469.

Strack, F., Stepper, S., & Martin, L. L. (1988). "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.

Strongman, K., . T. (2003). The psychology of Emotion, John WIley and Sons Ltd.

Strongman, K. T. (2003). The Psychology of Emotion: From Everyday Life to Theory, Wiley.

Sun-Microsystems. (2009). "My SQL Home Page." from http://www.mysql.com/.

Talbot-Smith, M. (2001). Audio Engineer's Reference Book, Second Edition, Focal Press

Tatham, M. and C. Morton (2004). Expression in speech. New York, Oxford University Press.

Thambirajah, M. S. (2004). Psychological Basis of Psychiatry, Churchill Livingstone

Theodoros, G. (2009). A dimensional approach to emotion recognition of speech from movies.

Titze, I. R. (1994). Principles of voice production, Prentice Hall.

Tolkmitt, F. J. and K. Scherer (1986). "Effect of Experimentally Induced Stress on Vocal Parameters." Journal of Experimental Psychology: Human Perception and Performance **12**(3): 302-313.

Tran, V. (2004). The Influence of Emotions on Decision-Making Processes in Management Teams. Psychology. Geneva, University of Geneve. **Ph.D:** 219.

Trouvain, J. and J. W. Barry (2000). The Prosody of Excitement in Horse Racing Commentaries. ITRW on Speech and Emotion. Newcastle, Northen Ireland, UK**:** 6.

Tsao, Y.-C., G. Weismer, et al. (2006). "The effect of intertalker speech rate variation on acoustic vowel space." The Journal of the Acoustical Society of America **119**(2): 1074-1082.

Twitter. (2010). "Twitter About Page." from http://twitter.com/about.

Vainio, M., and Altosaar, T. (1998). "Modeling The Microprosody of Pitch and Loudness For Speech Synthesis With Neural Networks."

Velten, E. (1968). "A laboratory task for induction of mood states." Behaviour Research and Therapy **6**: 473-482.

Ververidis, D., and Kotropoulos, C. (2006). "Emotional speech recognition: Resources, features, and methods " Speech Communication **48**(9): 1162-1181.

Ververidis, D. and C. Kotropoulos (2006). "Emotional speech recognition : Resources, features, and methods." Speech communication **48**(9): 1162-1181.

Watson, J. B. (1929). Psychology. From the Standpoint of a Behaviorist. Philadelphia, Lippincott.

Weiss, F., Blum, G.S., Gleberman, L. (1987). "Anatomically based measurement of facial expressions in simulated versus hypnotically induced effect." Motivation and Emotion **11**(67-81).

Weizenbaum., J. (1976). Computer Power and Human Reason. San Francisco Freeman.

Wendt, B., & Scheich, H. (2002). The Magdeburger Prosodie-Korpus. Speech Prosody Conference, Aix-en-Provence, France.

Werner, S. and E. Keller (1994). Prosodic Aspects of Speech. Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of tha Art, and Future Challenges. E. Keller. Chichester, John Wiley**:** 23-40.

Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). "Relative effectiveness and validity of mood induction procedures: a meta analysis." European Journal of Social Psychology **26**: 557-580.

Wichmann, A. (2000). The attitudinal effects of prosody and how they relate to emotion. ISCA Workshop on Speech and Emotion, , Belfast.

Wittenburg, P., W. Peters, et al. (2002). Metadata proposals for corpora and lexica. The 3rd International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas, Canary Islands**:** 5.

Worth, L. T., Mackie, D. M. (1987). "Cognitive mediation of positive affect in persuasion." Social Cognition **5**: 76-94.

Wundt, W. (1896). Grundriss der Psychologie. Leipzig. , Verlag von Wilhelm Engelmann.

XML, W. W. W. C. (2008). "Extensible Markup Language." from http://www.w3.org/XML/.

Yang, Y.-H., Y.-C. Lin, et al. (2009). Personalised Music Emotion Recognition. Annual ACM Conference on Research and Development in Information Retrieval, Boston, MA, USA.

Appendix A.Tables Of Primary Emotions Across Researchers

Lazarus (1999)	Ekman (1999)	Buck (1999)	Lewis & Harvard (1993)	Banse & Scherer (1996)	Cowie et al. (1999)
Anger	Anger	Anger	Anger	Rage	Anger
Fright	Fear	Fear	Fear	Fear	Afraid
Sadness	Sadness	Sadness	Sadness	Sadness	Sad
Anxiety		Anxiety	Anxiety	Anxiety	Worried
Happiness	Pleasure	Happiness	Happiness	Happiness	Happy
	Amusement		Humour		Amused
	Satisfaction				Pleased
	Contentment				Content
		Interested			
		Curious			
		Surprised			
	Excitement				Excited
		Bored		Boredom	Bored
					Relaxed
		Burnt-out			
Disgust	Disgust	Disgust	Disgust	Disgust	
	Contempt	Scorn		Scorn	
Pride	Pride	Pride	Pride		
		Arrogance			
Jealousy		Jealousy			
Envy		Envy			
Shame	Shame	Shame	Shame	Shame/Guilt	
Guilt	Guilt	Guilt	Guilt		
	Embarrassment		Embarrassment		Disappointed
					Confident
Love			Love		Loving
					Affectionate
Compassion		Pity			
		Moral Rapture			
		Moral Indignation			
Aesthetic					

Arnold (1960)	Ekman, Friesen & Ellsworth (1982)	Frijda (1986)	Gray (1982)	Izard (1871)	James (1884)	McDougall (1926)	Mower (1960)	Oatley & Johnson-Laird (1987)	Panskepp (1982)	Plutchik (1980)	Tomkins (1984)	Watson (1930)	Weiner & Graham (1984)
Anger	Anger			Anger	Rage	Anger		Anger	Rage	Anger	Anger	Rage	
Fear	Fear		Rage/Terror	Fear	Fear	Fear			Fear	Fear	Fear	Fear	
Sadness	Sadness	Sorrow			Grief			Sadness		Sadness			Sadness
Despair			Anxiety	Distress				Anxiety	Panic		Distress		
Hope	Joy	Happiness	Joy	Joy		Elation	Pleasure	Happiness		Joy	Joy		Happiness
										Acceptance			
							Pain						
						Subjection							
Courage													
Desire		Desire											
Aversion													
		Wonder				Wonder							
		Interest		Interest							Interest		
	Surprise	Surprise		Surprise							Surprise		
	Disgust			Disgust		Disgust		Disgust		Disgust	Disgust		
Hate				Contempt							Contempt		
				Shame							Shame		
				Guilt									
Dejection													
Love					Love	Tender-emotion						Love	

Appendix B. Action Script 3 Code From The Listening Tool

```
<?xml version="1.0" encoding="utf-8"?>
<mx:Application xmlns:mx="http://www.adobe.com/2006/mxml" layout="absolute"
xmlns:ns1="*" alpha="1.0" borderStyle="none" horizontalAlign="center"
verticalAlign="middle" creationComplete="init();" dropShadowEnabled="false"
backgroundGradientAlphas="0" backgroundGradientColors="0"
backgroundColor="#404040" dropShadowColor="#FFFFFF">
<mx:states>
<mx:State name="Finished">
<mx:AddChild relativeTo="{canvas1}position="lastChild">
<mx:Text x="96" y="199.45" text="Thank you for partaking. You may now close
this application. " width="415" fontSize="23" color="#020202"
textAlign="center" id="text6"/>
</mx:AddChild>
<mx:SetProperty target="{text5}" name="x" value="125.5"/>
<mx:SetProperty target="{text5}" name="y" value="105.9"/>
<mx:SetProperty target="{image1}" name="x"/>
<mx:SetProperty target="{image1}" name="y" value="10"/>
<mx:SetProperty target="{canvas1}" name="width" value="606"/>
<mx:SetProperty target="{canvas1}" name="height" value="296"/>
<mx:SetStyle target="{canvas1}" name="verticalCenter"/>
<mx:SetProperty target="{canvas2}(Ekman 1993)" name="width" value="607"/>
<mx:SetStyle target="{image1}" name="horizontalCenter" value="0"/>
<mx:SetProperty target="{canvas2}" name="y" value="0"/>
<mx:SetProperty target="{canvas1}" name="y" value="107"/>
<mx:SetProperty target="{canvas1}" name="x" value="174"/>
<mx:SetStyle target="{canvas1}" name="horizontalCenter"/>

</mx:State>
</mx:states>
<mx:Script>
<![CDATA[
import mx.rpc.events.AbstractEvent;

// these variables relate to the the sound playing abilities of the tool.
import flash.media.Sound;
import flash.media.SoundChannel;
import flash.net.URLRequest;
// Blur window effect import

import mx.controls.Alert;
           import mx.events.CloseEvent;


import mx.events.SliderEvent;
import mx.controls.sliderClasses.Slider;

//Blur popup window import
  import mx.controls.Alert;
         import mx.events.CloseEvent;
```

```
private var PlayPress:String;
private var songName:String;
private var sound:Sound;
private var channel:SoundChannel;
private var Mp3App:String;


import mx.rpc.soap.WebService;
import mx.rpc.events.ResultEvent;

private var corpus:WebService;
private var random:Object;
private var sounds:Array;
private var randomArray:Object;
private var rateNow1:Boolean = false;
private var rateNow2:Boolean = false;
private var index:Number = 0;

// these two functions get the value of the sliders and populate a text box
with them. Can then be asigned to a new variable and written to a database.

private function init():void {
corpus = new WebService();
sounds = new Array();
corpus.wsdl = "http://corpus.dmc.dit.ie/ranks/wsdl";
corpus.loadWSDL();
corpus.RandomArray.addEventListener("result", resultHandler);
corpus.RandomArray();
}

private function resultHandler(event:ResultEvent):void {
randomArray = event.result;
var count:Number = 0;
while (count < randomArray.length){
sounds.push(new String("http://corpus.dmc.dit.ie/mp3/" +
randomArray[count] + ".mp3"));
count++;
}
}


private function sliderChange(event:SliderEvent):void {
        var currentSlider:Slider=Slider(event.currentTarget);
        rateNow1 = true;
        if(rateNow1 == true){
        if(rateNow2 == true){
        //rateBtn.enabled = true;
        //NotrateBtn.enabled = true;
        }
        }
}


private function sliderChange2(event:SliderEvent):void {
        var currentSlider2:Slider=Slider(event.currentTarget);

        rateNow2 = true;
```

```
        if(rateNow1 == true){
        if(rateNow2 == true){
        //rateBtn.enabled = true;
        //NotrateBtn.enabled = true;
        }
        }
}


private function startPlay(replay:Boolean):void {

var urlRef:String = sounds[index];
sound = new Sound(new URLRequest(urlRef));

if(replay == true){
channel.stop();

//This makes sure the play buton is pressed before the rate and do not rate
buttons become active

}


//Made changes to the way the play button functions. Once play has been
pressed the rating buttons become active.
rateBtn.enabled = true;
        NotrateBtn.enabled = true;
//Play
channel = sound.play();

//Play button invisible
playBtn.visible = false;
playBtn.enabled = false;

//Replay visible
rePlayBtn.visible = true;
rePlayBtn.enabled = true;

}


private function rate(rate:Boolean):void {

playBtn.visible = true;
playBtn.enabled = true;

rePlayBtn.visible = false;
rePlayBtn.enabled = false;

rateBtn.enabled = false;
NotrateBtn.enabled = false;

rateNow1 = false;
rateNow2 = false;

if(rate == true){
corpus.New(randomArray[index], Active_Passive.value,
Postive_Negative.value);
```

```
}else{
corpus.New(randomArray[index], 999, 999);
}


//Increment the index to play the next clip
index++;

if(index > 0){
clip1.enabled = false;
}
if(index > 1){
clip2.enabled = false;
}
if(index > 2){
clip3.enabled = false;
}
if(index > 3){
clip4.enabled = false;
}
if(index > 4){
clip5.enabled = false;
}
if(index > 5){
clip6.enabled = false;
}
if(index > 6){
clip7.enabled = false;
}
if(index > 7){
clip8.enabled = false;
}
if(index > 8){
clip9.enabled = false;
}
if(index > 9){
clip10.enabled = false;
}

if(index == 10){


//Alert.show('You have rated 10 clips, would you like to rate 10 more?',
'Message');
 Alert.show("You have rated 10 clips, would you like to rate 10
more?", "Finished!", 3, this, alertClickHandler);
}



channel.stop();

//Reset the Slider values
Active_Passive.value = 5;
Postive_Negative.value = 5;

lab.text = "Audio Clip " + (index + 1);
```

```
}


            // Event handler function for displaying the selected Alert
button.
            private function alertClickHandler(event:CloseEvent):void {
               if (event.detail==Alert.YES)
           reloadpage();
               else
                 currentState = "Finished";
            }



//java to reload the page *************
private function reloadpage():void
{
var ref:URLRequest = new
URLRequest("javascript:location.reload(true)");
navigateToURL(ref, "_self");
}


//************************************

]]>
</mx:Script>
<mx:Canvas width="750" height="479" id="canvas1" horizontalCenter="0"
verticalCenter="-36" backgroundColor="#201f20" backgroundAlpha="0.0"
borderStyle="solid" cornerRadius="0" resizeEffect="Resize"
borderThickness="0">

<mx:Canvas x="0" y="-1" width="750" height="114.95" borderStyle="solid"
cornerRadius="0" alpha="1.0" themeColor="#000000" backgroundColor="#a62023"
id="canvas2" borderThickness="0" backgroundAlpha="1.0">
<mx:Image x="20" y="7.95" source="MMILLogo4.png" id="image1"/>
<mx:Text x="435" y="43.9" text="Emotional Rating Tool" width="270"
height="49.402557" fontSize="20" color="#ffffff" id="text5"/>
</mx:Canvas>



<mx:Canvas x="0" y="112.05" width="367" height="336.95" borderStyle="solid"
cornerRadius="0" alpha="1.0" themeColor="#000000" backgroundColor="#d0d0d0"
id="canvas0" borderThickness="0">
<mx:Label x="33" y="23" id="lab" text="AUDIO CLIP ONE" width="157"
height="26.9" fontSize="11" color="#020202" fontFamily="Verdana"
fontWeight="bold"/>
<mx:SWFLoader x="33" y="67.9" source="bvsoundSpectrum.swf" width="71"
height="121" id="swfloader3"/>
<mx:SWFLoader x="33" y="67.9" source="bvsoundSpectrum2.swf" width="71"
height="121" autoLoad="true" scaleContent="true" id="swfloader2"/>
<mx:Button label="Play" color="#FFFFFF" width="58" fontSize="10"
cornerRadius="13" id="playBtn"  click="startPlay(false);"  x="22.5"
y="219.9" alpha="1.0" fillAlphas="[1.0, 1.0]" fillColors="[#aa1d21,
#9e1b1f]" borderColor="#9b2d1a" themeColor="#A8311C" height="23"/>
<mx:Button x="22.5" y="286.95" label="1" id="clip1" cornerRadius="0"
borderColor="#D0D0D0" color="#ffffff" width="25" height="25"
```

```
fillAlphas="[1.0, 1.0]" themeColor="#b3351e" fillColors="[#000000,
#000000]"/>
<mx:Button x="55.5" y="286.95" label="2" id="clip2"  cornerRadius="0"
borderColor="#D0D0D0" color="#ffffff" width="25" height="25"
fillAlphas="[1.0, 1.0]" themeColor="#b3351e" fillColors="[#000000,
#000000]"/>
<mx:Button x="121.5" y="286.95" label="4" id="clip4"    cornerRadius="0"
borderColor="#D0D0D0" color="#ffffff" width="25" height="25"
fillAlphas="[1.0, 1.0]" themeColor="#b3351e" fillColors="[#000000,
#000000]"/>
<mx:Button x="88.5" y="286.95" label="3" id="clip3"   cornerRadius="0"
borderColor="#D0D0D0" color="#ffffff" width="25" height="25"
fillAlphas="[1.0, 1.0]" themeColor="#b3351e" fillColors="[#000000,
#000000]"/>
<mx:Button x="154.5" y="286.95" label="5" id="clip5"   cornerRadius="0"
borderColor="#D0D0D0" color="#ffffff" width="25" height="25"
fillAlphas="[1.0, 1.0]" themeColor="#b3351e" fillColors="[#000000,
#000000]"/>
<mx:Button x="187.5" y="286.95" label="6" id="clip6"   cornerRadius="0"
borderColor="#D0D0D0" color="#ffffff" width="25" height="25"
fillAlphas="[1.0, 1.0]" themeColor="#b3351e" fillColors="[#000000,
#000000]"/>
<mx:Button x="220.5" y="286.95" label="7" id="clip7"  cornerRadius="0"
borderColor="#D0D0D0" color="#ffffff" width="25" height="25"
fillAlphas="[1.0, 1.0]" themeColor="#b3351e" fillColors="[#000000,
#000000]"/>
<mx:Button x="253.5" y="286.95" label="8" id="clip8"   cornerRadius="0"
borderColor="#D0D0D0" color="#ffffff" width="25" height="25"
fillAlphas="[1.0, 1.0]" themeColor="#b3351e" fillColors="[#000000,
#000000]"/>
<mx:Button x="286.5" y="286.95" label="9" id="clip9"  cornerRadius="0"
borderColor="#D0D0D0" color="#ffffff" width="25" height="25"
fillAlphas="[1.0, 1.0]" themeColor="#b3351e" fillColors="[#000000,
#000000]"/>
<mx:Button x="319.5" y="286.95" label="10" id="clip10"   cornerRadius="0"
borderColor="#D0D0D0" color="#ffffff" width="25" height="25"
fillAlphas="[1.0, 1.0]" themeColor="#b3351e" fillColors="[#000000,
#000000]"/>
</mx:Canvas>



<mx:Canvas x="366" y="112.05" width="384" height="336.95"
borderStyle="solid" cornerRadius="0" alpha="1.0" themeColor="#000000"
backgroundColor="#ffffff" id="canvas0" borderThickness="0">
<mx:HSlider x="62.5" y="57.95" id="Active_Passive" minimum="0" maximum="10"
value = "5" enabled="true" snapInterval=".5" change="sliderChange(event);"
width="256"/>
<mx:HSlider x="62.5" y="137.95" id="Postive_Negative" minimum="0"
maximum="10" value= "5" snapInterval=".5" change="sliderChange2(event);"
width="256"/>
<mx:Text x="267.5" y="26.95" text="Active" width="48" id="text1"
color="#000000" fontSize="12" fontFamily="Verdana" textAlign="right"/>
<mx:Text x="67.5" y="106.95" text="Negative" id="text2" color="#000000"
fontSize="12" fontFamily="Verdana" width="69"/>
<mx:Text x="67.5" y="26.95" text="Passive" color="#000000" fontSize="12"
id="text4" fontFamily="Verdana" width="69"/>
```

```
<mx:Text x="267.5" y="106.95" text="Positive" color="#000000" fontSize="12"
id="text3" fontFamily="Verdana"/>
<mx:ApplicationControlBar x="59.5" y="257.95" width="263" dock="false"
cornerRadius="0" id="applicationcontrolbar1" fillColors="[#f1fffd, #fffefb]"
fillAlphas="[0.0, 0.0]" height="38" color="#ffffff" alpha="1.0">
</mx:ApplicationControlBar>
<mx:Button label="Do not rate" enabled="false" color="#ffffff" width="103"
fontSize="12" id="NotrateBtn"  cornerRadius="15" click="rate(false);"
alpha="1.0" fillAlphas="[1.0, 1.0]" fillColors="[#7d7d7d, #7e7e7e]"
x="155.5" y="281.95" height="24"/>
<mx:Button label="Rate" enabled="false" color="#ffffff" width="80"
fontSize="12" id="rateBtn"  cornerRadius="15" click="rate(true);"
fillAlphas="[1.0, 1.0]" fillColors="[#000000, #000000]"
borderColor="#ffffff" height="24" alpha="1.0" x="67.5" y="281.95"/>

</mx:Canvas>
<mx:Button label="Replay" color="#030303" width="80" fontSize="12"
cornerRadius="15" id="rePlayBtn"  click="startPlay(true);" visible="false"
x="134" y="352"/>

</mx:Canvas>

</mx:Application>
```

Appendix C. Consent Forms For Participants Taking Part In All MIP Experiments

digital media centre

Research Participant Release Form

Participant Name (in print):

_____

_____

You have been asked to be a participant in ongoing speech research within the Cognition Speech and Audio Lab (CSAL) which is part of the Digital Media Centre (DMC) in the Dublin Institute of Technology (DIT). The team of researchers within the CSAL may want to use the recording which has been made of your voice for future research purposes and the collected data and findings may possibly feature in future research publications. As well as this the recording of your voice may also form part of a Speech Database/Corpus that will be used for further research by the CSAL and other academic third parties.

The CSAL research team will endeavour the following:

- To protect the welfare and dignity of the participant.
- To respect the individual's freedom to decline participation.
- To maintain confidentiality of the research data.
- To be responsible for maintaining ethical standards.
- To NOT specifically identify individuals with their data unless it is necessary, and then only after the individual has given consent.
- To take every precaution and make every effort to minimise potential risk to participants.
- To only use the data supplied by the participant with their full consent.

I herby give my consent to the CSAL research team in the DIT to use my voice recording and any data supplied by me on the accompanying form for research purposes and possible further publications:

Name (block capitals): _____

Date: _____

Signature: _____

digital media centre

Research Participant Additional Information Form

This form consists of brief questions that may be of relevance to the research being undertaken using your voice recordings. You can fill out as many or as little as you chose.

Date of Birth:

Gender:

Nationality:

First Language:

Please list any other languages that you are fluent or proficient in:

_____

_____

_____

Do you suffer from any speech impediments? Yes ☐ No ☐

Have you ever taken part in similar research before? Yes ☐ No ☐

Details: _____

_____

_____

_____

Appendix D. Table Of Ratings Received For The Shipwreck MIP

Rating_Count					
Count		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7.00	2	1.1	1.1	1.1
	8.00	1	.6	.6	1.7
	10.00	2	1.1	1.1	2.8
	11.00	4	2.3	2.3	5.1
	12.00	2	1.1	1.1	6.2
	13.00	11	6.2	6.2	12.4
	14.00	12	6.8	6.8	19.2
	15.00	8	4.5	4.5	23.7
	16.00	10	5.6	5.6	29.4
	17.00	12	6.8	6.8	36.2
	18.00	11	6.2	6.2	42.4
	19.00	14	7.9	7.9	50.3
	20.00	9	5.1	5.1	55.4
	21.00	11	6.2	6.2	61.6
	22.00	11	6.2	6.2	67.8
	23.00	7	4.0	4.0	71.8
	24.00	6	3.4	3.4	75.1
	25.00	4	2.3	2.3	77.4
	26.00	8	4.5	4.5	81.9
	27.00	5	2.8	2.8	84.7
	28.00	7	4.0	4.0	88.7
	29.00	2	1.1	1.1	89.8
	30.00	2	1.1	1.1	91.0
	31.00	1	.6	.6	91.5
	33.00	1	.6	.6	92.1
	35.00	3	1.7	1.7	93.8
	36.00	2	1.1	1.1	94.9
	37.00	2	1.1	1.1	96.0
	42.00	2	1.1	1.1	97.2
	43.00	1	.6	.6	97.7
	44.00	2	1.1	1.1	98.9
	45.00	2	1.1	1.1	100.0
	Total	177	100.0	100.0	

Appendix E. Table Of Median And IQR Values For All Assets On The Activation Dimension

Audio_ID	Median	25th Percentile	75th Percentile	IQR
138	5	5	5	0
69	8	7	8	1
77	6.75	6	7	1
78	6.5	6	7	1
97	6.5	6	7	1
98	7.5	7	8	1
106	8	7.5	8.5	1
107	7.5	7	8	1
169	4	3	4	1
203	6.5	6	7	1
221	4	3	4	1
223	6	6	7	1
225	7	6.5	7.5	1
226	8	7.25	8.25	1
40	3	2	3.5	1.5
46	6.5	5	6.5	1.5
53	6	5.5	7	1.5
72	7.5	6.5	8	1.5
92	7	6.5	8	1.5
95	7	6	7.5	1.5
103	6.5	5.5	7	1.5
117	6	5.25	6.75	1.5
144	7	6	7.5	1.5
157	6.5	6	7.5	1.5
158	7.25	6.75	8.25	1.5
162	6.75	6	7.5	1.5
175	7	5.5	7	1.5
184	6	5	6.5	1.5
185	6.5	5.25	6.75	1.5
190	7.5	6	7.5	1.5
197	7	6	7.5	1.5
206	7.25	6.5	8	1.5
210	6.5	6	7.5	1.5
238	6.5	5	6.5	1.5
240	6.5	5.5	7	1.5
241	7.5	6.5	8	1.5
108	7.75	6.75	8.5	1.75
132	7	5.75	7.5	1.75
174	5.5	4.5	6.25	1.75
31	5.5	4.5	6.5	2
41	5.5	4.5	6.5	2
50	7	6	8	2
51	6	4.75	6.75	2
52	6.5	5.5	7.5	2
63	5	3.5	5.5	2
73	6	5	7	2
80	7	5.5	7.5	2
83	3.5	2.75	4.75	2
85	6	5	7	2
93	4	3.5	5.5	2
100	3	1.5	3.5	2
105	5.25	4	6	2
112	4	3.5	5.5	2
114	6	5	7	2
115	6.5	5.5	7.5	2
119	6	5	7	2

125	6.5	5	7	2
130	4	3	5	2
131	6	5.5	7.5	2
133	6.5	6	8	2
135	4	4	6	2
141	6.75	6	8	2
154	8.5	7	9	2
160	5	4	6	2
170	4	3.5	5.5	2
171	2.5	1.5	3.5	2
181	3.5	2	4	2
189	6.5	5.5	7.5	2
192	6.5	5	7	2
195	6	5	7	2
205	7.5	6	8	2
214	6.5	5.5	7.5	2
217	6.25	5	7	2
219	6	5	7	2
227	8.5	7	9	2
228	6.5	5	7	2
79	6	5	7.25	2.25
177	6	4.25	6.5	2.25
208	5.75	4.5	6.75	2.25
235	3	1.75	4	2.25
33	5.25	3.75	6.25	2.5
34	4.5	3.5	6	2.5
43	5.5	4	6.5	2.5
47	4	3.5	6	2.5
61	4.5	3.5	6	2.5
65	4	3	5.5	2.5
66	5.5	4	6.5	2.5
68	4	2.5	5	2.5
75	6.5	5.5	8	2.5
84	7.5	6.5	9	2.5
94	7	5.75	8.25	2.5

96	5.5	4	6.5	2.5
118	3.5	2.5	5	2.5
121	6	5	7.5	2.5
136	6.25	5	7.5	2.5
156	6.75	5	7.5	2.5
167	6	4	6.5	2.5
173	3.25	3	5.5	2.5
176	7.25	6	8.5	2.5
178	5.5	4	6.5	2.5
187	7.5	5.5	8	2.5
198	6.5	5.5	8	2.5
199	3.5	2.5	5	2.5
200	4.25	3	5.5	2.5
215	4.5	3.5	6	2.5
218	5.25	3.5	6	2.5
222	5	3.5	6	2.5
87	5.75	4.5	7.25	2.75
104	4.5	3.25	6	2.75
153	4.5	3.25	6	2.75
196	6.5	4.75	7.5	2.75
30	5.75	4	7	3
42	4	3	6	3
54	5.75	3.5	6.5	3
64	3.5	2.5	5.5	3
70	7	6	9	3
71	7	5	8	3
76	5.5	4	7	3
81	6	4	7	3
82	6	4	7	3
89	3.5	2.5	5.5	3
99	5.75	4	7	3
120	5.75	4	7	3
129	6	4	7	3
139	6.5	4.5	7.5	3
155	5	3.5	6.5	3

159	6.25	4	7	3
168	6	4	7	3
179	2	1	4	3
193	4	3	6	3
194	5.5	4	7	3
204	6.5	4.5	7.5	3
207	5	3.5	6.5	3
211	3.5	2.5	5.5	3
213	5	3	6	3
216	4.5	3	6	3
220	4.5	3	6	3
230	2.75	1.5	4.5	3
231	3.75	2.5	5.5	3
247	2.75	2.25	5.25	3
49	3.5	2.75	6	3.25
74	5.5	3.5	6.75	3.25
229	3.5	2	5.25	3.25
233	6	3.75	7	3.25
62	6.5	4	7.5	3.5
88	5	3	6.5	3.5
90	6	3.5	7	3.5
91	4	3	6.5	3.5
101	6.5	4.5	8	3.5
102	3	2	5.5	3.5
109	6.5	5	8.5	3.5
110	5	3.5	7	3.5
111	5.5	2.5	6	3.5
113	5.5	3.5	7	3.5
116	6	3.5	7	3.5
137	5	2.5	6	3.5
165	6.5	3.5	7	3.5
172	3	2	5.5	3.5
180	3.5	2.5	6	3.5
186	6	4	7.5	3.5
191	6	3.5	7	3.5
201	3.5	2	5.5	3.5
202	3.25	2	5.5	3.5
212	6	4	7.5	3.5
232	3	2	5.5	3.5
236	2	1.5	5	3.5
242	4.25	2.5	6	3.5
134	7	4	8	4
182	3	2	6	4
209	4	2.5	6.5	4
224	6.5	3.5	7.5	4
237	3.25	2	6.25	4.25
48	4	2	6.5	4.5
188	7	4	8.5	4.5
239	6.5	3	7.5	4.5
248	4	3	7.5	4.5
183	4.5	1.5	6.5	5

Appendix F. Median And IQR Values For All Assets on the Evaluation Dimension

Audio_ID	Median	25th Percentile	75th Percentile	IQR
138	5	5	5	0
181	3.5	3.5	4	0.5
190	3.5	3	3.5	0.5
47	5	5	6	1
50	6	5.5	6.5	1
63	5.5	5	6	1
105	5.5	5	6	1
112	4.5	4	5	1
120	6.5	6	7	1
134	6	5.5	6.5	1
184	4.5	4.5	5.5	1
193	4	4	5	1
201	3	2	3	1
202	5	5	6	1
213	3.5	3	4	1
219	5.5	5	6	1
226	2.5	2.5	3.5	1
240	3.5	3	4	1
31	5.5	4.75	6	1.25
72	7.5	6.75	8	1.25
104	6.25	6	7.25	1.25
30	6	5	6.5	1.5
46	6	5	6.5	1.5
61	5	4	5.5	1.5
64	3.75	3.5	5	1.5
69	7.5	7	8.5	1.5
74	6	5.75	7.25	1.5
82	6.5	6	7.5	1.5
96	6	5	6.5	1.5
107	7.5	6.5	8	1.5
125	5	4	5.5	1.5
135	4	4	5.5	1.5
137	5.25	5	6.5	1.5
160	5	4	5.5	1.5
169	2	1.5	3	1.5
170	4	3	4.5	1.5
172	4.5	4	5.5	1.5
179	2.75	2	3.5	1.5
196	4.5	3.75	5.25	1.5
200	5.25	5	6.5	1.5
203	4	3.5	5	1.5
209	2.25	1.75	3.25	1.5
211	4.5	3.5	5	1.5
221	5	4.5	6	1.5
222	4.5	4	5.5	1.5
225	4	3	4.5	1.5
227	2.5	2	3.5	1.5
238	5	4.5	6	1.5
242	3	2.5	4	1.5
43	5	3.75	5.5	1.75
51	6.25	5.5	7.25	1.75
237	4.25	3.25	5	1.75
34	4	4	6	2
40	3.5	2.5	4.5	2

42	4.5	3.5	5.5	2
52	6	5	7	2
53	5	4	6	2
65	5	4.5	6.5	2
66	6	4.5	6.5	2
68	4.75	3	5	2
70	8.5	7.5	9.5	2
81	6	5	7	2
89	5.5	4.5	6.5	2
92	6	5	7	2
93	4	3.5	5.5	2
98	8	7	9	2
99	5.5	4	6	2
101	6	5	7	2
102	4	3	5	2
106	7	5.5	7.5	2
115	5	4.5	6.5	2
130	4.5	3.5	5.5	2
139	6.5	5.5	7.5	2
153	3	1.75	3.75	2
156	5	3.5	5.5	2
162	6.5	5.5	7.5	2
171	2.5	2	4	2
178	5	4.5	6.5	2
182	3	2	4	2
185	5.5	4.5	6.5	2
188	3	2	4	2
194	5.5	4	6	2
197	6	5	7	2
199	5	3.75	5.75	2
204	4.5	4	6	2
210	3.5	2.5	4.5	2
217	5.5	4.75	6.75	2
223	5	4	6	2
236	4.5	3	5	2

49	5	3.75	6	2.25
80	8	6.25	8.5	2.25
83	3	2	4.25	2.25
174	4.25	3.5	5.75	2.25
229	4.5	3	5.25	2.25
33	5	3.5	6	2.5
41	5	3.5	6	2.5
62	5	4	6.5	2.5
76	5.5	4.5	7	2.5
79	6.75	5.75	8.25	2.5
90	7	5.5	8	2.5
109	7	6	8.5	2.5
110	3.75	3.5	6	2.5
111	4	2.5	5	2.5
114	6	4.5	7	2.5
117	6	4	6.5	2.5
118	3.5	1.5	4	2.5
119	5	4	6.5	2.5
121	5.5	4.5	7	2.5
131	7	5	7.5	2.5
133	7.25	6	8.5	2.5
155	4.5	3.5	6	2.5
157	6	5	7.5	2.5
159	5.75	4	6.5	2.5
167	5.5	4	6.5	2.5
173	4	2.5	5	2.5
189	5.5	4	6.5	2.5
191	5	3	5.5	2.5
212	5	3	5.5	2.5
215	3.75	3	5.5	2.5
232	5.5	4.5	7	2.5
241	3.5	2.5	5	2.5
248	3	2	4.5	2.5
233	5	4	6.75	2.75
77	5.75	4	7	3

78	3.5	2.5	5.5	3
95	6.5	5	8	3
97	5.5	4.5	7.5	3
100	4	2.5	5.5	3
113	4.5	3	6	3
129	4.75	3.5	6.5	3
132	6.75	4.75	7.75	3
144	6	3.5	6.5	3
168	5.5	4	7	3
175	4.5	3	6	3
176	6.5	5	8	3
177	5.5	4	7	3
187	3	2.5	5.5	3
216	4	2.5	5.5	3
220	4	3	6	3
231	4	3	6	3
247	2	1.5	4.5	3
54	4.75	2.75	6	3.25
88	4	2.5	5.75	3.25
205	2	1.5	4.75	3.25
235	3.75	1.75	5	3.25
48	5	3.5	7	3.5
71	7.5	5.5	9	3.5
73	5	4	7.5	3.5
75	4	2.5	6	3.5
85	6.5	4.5	8	3.5
87	4.25	3.25	6.75	3.5
91	6.25	3.5	7	3.5
103	5	3.5	7	3.5
116	5.5	3.5	7	3.5
180	3.5	2.5	6	3.5
183	2.5	0.5	4	3.5
186	3.25	2.5	6	3.5
192	6	4	7.5	3.5
195	5	2.5	6	3.5
206	2.5	0	3.5	3.5
207	3.75	2.5	6	3.5
208	4.75	3.75	7.25	3.5
214	4.25	2.5	6	3.5
228	5	3	6.5	3.5
230	4.75	2.5	6	3.5
239	4	2.5	6	3.5
84	4.5	3	7	4
94	5.75	3.5	7.5	4
136	5.5	3.5	7.5	4
141	6	3	7	4
165	4	2.5	6.5	4
218	4	2.5	6.5	4
158	6	3.5	7.75	4.25
198	5	3	7.5	4.5
224	4.25	2	6.5	4.5
154	5.5	2	7	5
108	3.5	2.25	7.5	5.25

Appendix G. Cluster Analysis Tables For Assets With An IQR Less Than 2 On The Activation Dimension

Initial Cluster Centers for All Clustering Procedures

	Cluster																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Median	0	.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10

Iteration History[a]

Iteration	Change in Cluster Centers																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	.	.	.	.	.	.	.000	.	.000	.	.00	.00	.00	.042	.056	.042	.00	.	.	.	.
2	.	.	.	.	.	.	.00	.	.00	.	.00	.00	.00	.00	.00	.00	.00	.	.	.	.

Final Cluster Centers

	Cluster																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Median	.	.	.	.	.	.	3.00	.	4.00	.	5.00	5.50	6.00	6.54	7.06	7.54	8.00	.	.	.	.

Number of Cases in each Cluster		
Cluster	1	.000
	2	.000
	3	.000
	4	.000
	5	.000
	6	.000
	7	1.000
	8	.000
	9	2.000
	10	.000
	11	1.000
	12	1.000
	13	4.000
	14	12.000
	15	9.000
	16	6.000
	17	3.000
	18	.000
	19	.000
	20	.000
	21	.000
Valid		39.000
Missing		.000

Cluster Membership			
Case Number	Audio_ID	Cluster	Distance
1	40	7	.000
2	46	14	.042
3	53	13	.000
4	69	17	.000
5	72	16	.042
6	77	14	.208
7	78	14	.042
8	92	15	.056
9	95	15	.056
10	97	14	.042
11	98	16	.042
12	103	14	.042
13	106	17	.000
14	107	16	.042
15	108	16	.208
16	117	13	.000
17	132	15	.056
18	138	11	.000
19	144	15	.056
20	157	14	.042
21	158	15	.194
22	162	14	.208
23	169	9	.000
24	174	12	.000
25	175	15	.056
26	184	13	.000
27	185	14	.042
28	190	16	.042
29	197	15	.056
30	203	14	.042
31	206	15	.194
32	210	14	.042
33	221	9	.000
34	223	13	.000
35	225	15	.056
36	226	17	.000
37	238	14	.042
38	240	14	.042
39	241	16	.042

Distances between Final Cluster Centers

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
7	.	.	.	.	.	.		.	1.000	.	2.000	2.500	3.000	3.542	4.056	4.542	5.000	.	.	.	.
8	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	1.000	.		.	1.000	1.500	2.000	2.542	3.056	3.542	4.000	.	.	.	.
10	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	2.000	.	1.000	.		.500	1.000	1.542	2.056	2.542	3.000	.	.	.	.
12	.	.	.	.	.	.	2.500	.	1.500	.	.500		.500	1.042	1.556	2.042	2.500	.	.	.	.
13	.	.	.	.	.	.	3.000	.	2.000	.	1.000	.500		.542	1.056	1.542	2.000	.	.	.	.
14	.	.	.	.	.	.	3.542	.	2.542	.	1.542	1.042	.542		.514	1.000	1.458	.	.	.	.
15	.	.	.	.	.	.	4.056	.	3.056	.	2.056	1.556	1.056	.514		.486	.944	.	.	.	.
16	.	.	.	.	.	.	4.542	.	3.542	.	2.542	2.042	1.542	1.000	.486		.458	.	.	.	.
17	.	.	.	.	.	.	5.000	.	4.000	.	3.000	2.500	2.000	1.458	.944	.458		.	.	.	.
18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.
19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.
20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.
21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	

Appendix H. Cluster Analysis Tables For Assets With An IQR Less Than 2 On The Evaluation Dimension

Initial Cluster Centers for All Clustering Procedures

	Cluster																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Median	0	.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10

Iteration History[a]

Iteration	Change in Cluster Centers																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	.	.	.	.	.125	.083	.000	.050	.042	.000	.045	.000	.063	.000	.	.000	.	.	.	.	.
2	.	.	.	.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.	.000	.	.	.	.	.

Final Cluster Centers

	Cluster																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Median				.	2.13	2.58	3	3.55	4.04	4.5	5.05	5.5	6.06	6.5	.	7.5	.				

Number of Cases in each Cluster		
Cluster	1	.000
	2	.000
	3	.000
	4	.000
	5	2.000
	6	3.000
	7	2.000
	8	5.000
	9	6.000
	10	6.000
	11	11.000
	12	4.000
	13	8.000
	14	2.000
	15	.000
	16	3.000
	17	.000
	18	.000
	19	.000
	20	.000
	21	.000
Valid		52.000
Missing		.000

Cluster Membership			
Case Number	Audio_ID	Cluster	Distance
1	30	13	.063
2	31	12	.000
3	43	11	.045
4	46	13	.063
5	47	11	.045
6	50	13	.063
7	51	13	.188
8	61	11	.045
9	63	12	.000
10	64	8	.200
11	69	16	.000
12	72	16	.000
13	74	13	.063
14	82	14	.000
15	96	13	.063
16	104	13	.188
17	105	12	.000
18	107	16	.000
19	112	10	.000
20	120	14	.000
21	125	11	.045
22	134	13	.063
23	135	9	.042
24	137	11	.205
25	138	11	.045
26	160	11	.045
27	169	5	.125
28	170	9	.042
29	172	10	.000
30	179	6	.167
31	181	8	.050
32	184	10	.000
33	190	8	.050
34	193	9	.042
35	196	10	.000
36	200	11	.205
37	201	7	.000
38	202	11	.045
39	203	9	.042
40	209	5	.125
41	211	10	.000
42	213	8	.050
43	219	12	.000
44	221	11	.045
45	222	10	.000
46	225	9	.042
47	226	6	.083
48	227	6	.083
49	237	9	.208
50	238	11	.045
51	240	8	.050
52	242	7	.000

Distances between Final Cluster Centers

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	.	.	.	.		.458	.875	1.425	1.917	2.375	2.920	3.375	3.938	4.375	.	5.375	.	.	.	.	.
6	.	.	.	.	.458		.417	.967	1.458	1.917	2.462	2.917	3.479	3.917	.	4.917	.	.	.	.	.
7	.	.	.	.	.875	.417		.550	1.042	1.500	2.045	2.500	3.063	3.500	.	4.500	.	.	.	.	.
8	.	.	.	.	1.425	.967	.550		.492	.950	1.495	1.950	2.513	2.950	.	3.950	.	.	.	.	.
9	.	.	.	.	1.917	1.458	1.042	.492		.458	1.004	1.458	2.021	2.458	.	3.458	.	.	.	.	.
10	.	.	.	.	2.375	1.917	1.500	.950	.458		.545	1.000	1.563	2.000	.	3.000	.	.	.	.	.
11	.	.	.	.	2.920	2.462	2.045	1.495	1.004	.545		.455	1.017	1.455	.	2.455	.	.	.	.	.
12	.	.	.	.	3.375	2.917	2.500	1.950	1.458	1.000	.455		.563	1.000	.	2.000	.	.	.	.	.
13	.	.	.	.	3.938	3.479	3.063	2.513	2.021	1.563	1.017	.563		.438	.	1.438	.	.	.	.	.
14	.	.	.	.	4.375	3.917	3.500	2.950	2.458	2.000	1.455	1.000	.438		.	1.000	.	.	.	.	.
15	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.
16	.	.	.	.	5.375	4.917	4.500	3.950	3.458	3.000	2.455	2.000	1.438	1.000	.		.	.	.	.	.
17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.
18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.
19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.
20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.
21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	

Appendix I. PRAAT Script Used For The Acoustic Analysis Of Emotional Speech Assets

Custom PRAAT script written to carry out a comprehensive batch analysis of emotional speech assets.

```
#-------------------------------------------------------------------------
--#
# This script was created by Brian Vaughan (brian.vaughan AT dit.ie)
#
# April 2010. Use it as you wish but if you do use it in a larger script of
#  #your own please acknowledge its source      #Its fairly simple &
automates a lot of analysis I needed      #
#for my PhD Research         #
#-------------------------------------------------------------------------
--#

#Loads all .wav sounds from directory

Create Strings as file list... list *.wav
numberOfFiles = Get number of strings

for ifile to numberOfFiles
   select Strings list
   fileName$ = Get string... ifile
   Read from file... 'fileName$'
endfor

#Loops through all the files loaded

i=2

while i< numberOfFiles+2
select i


#This runs the pitch analysis on the slected sound

mySound = selected("Sound")

#This gets the name of the sound file that is currently selected
object_name$ = selected$ ("Sound")

#This creates a directory with the file name appended to it
createDirectory ("Analysis_Files For Asset  'object_name$'")


#************************************************************************
**********

To Pitch... 0 75 600

#This gets the 25th and 7th quantiles as they ensure no pitch calculation
#errors due to large incorrect pitch calculations. Many thanks to Daniel
Hirst #on the PRAAT users group for this.

q1 = Get quantile... 0 0 0.25 Hertz
```

```
q3 = Get quantile... 0 0 0.75 Hertz
Remove
select mySound
To Pitch... 0 q1*0.75 q3*1.5
#*************************************************************************
**********
#This gets the autocorrelated pitch, better for intonation research

#mySound = selected("Sound")
#select mySound
#To Pitch... 0 75 600
#q1 = Get quantile... 0 0 0.25 Hertz
#q3 = Get quantile... 0 0 0.75 Hertz
#Remove
#select mySound
#To Pitch (ac)... 0 q1*0.75 15 no 0.03 0.45 0.01 0.35 0.14 q3*1.5

#This selects a decent size window in the drawing window and draws the pitch
contour to it
Select outer viewport... 0 12 0 6
Red
Erase all
Draw... 0 0 0 q3*1.5 yes
Marks left every... 1 20 yes yes yes
Marks bottom every... 1 0.1 yes yes yes
Draw inner box

#This writes the graphic to a pdf file
Write to PDF file... Analysis_Files For Asset
'object_name$'/Pitch_Contour_File 'object_name$'.pdf
Erase all

Draw logarithmic... 0 0 q1*0.75 q3*1.5 yes
Marks bottom every... 1 0.1 yes yes yes
Draw inner box
Write to PDF file... Analysis_Files For Asset
'object_name$'/Pitch_Hert(Log)Contour_File 'object_name$'.pdf
Erase all

Draw semitones... 0 0 -12 30 yes

Marks left every...  1 2.5 yes yes yes
Marks bottom every... 1 0.1 yes yes yes
Draw inner box
Write to PDF file... Analysis_Files For Asset
'object_name$'/Pitch_Contour_Semitones__File 'object_name$'.pdf
Erase all

#This section gets the min, max, mean, median and 25th and 75th quartile
values


pitchmin=Get minimum... 0 0 Hertz Parabolic
pitchmax=Get maximum... 0 0 Hertz Parabolic
pitchmean=Get mean... 0 0 Hertz
pitchstnddev=Get standard deviation... 0 0 Hertz
pitchmedian=Get quantile... 0 0 0.5 Hertz
pitch1stquant=Get quantile... 0 0 0.25 Hertz
pitch3rdquant=Get quantile... 0 0 0.75 Hertz
pitchrange=pitchmax-pitchmin
pitchiqr=pitch3rdquant-pitch1stquant

#Gets the pitch values as semitones and the hertz(Log) value
#This is equivilant to converting to semtiones and converting back to Hertz.

semitonemin=Get minimum... 0 0 "semitones re 1 Hz" Parabolic
semitonemax=Get maximum... 0 0 "semitones re 1 Hz" Parabolic
```

```
semitonemedian= Get quantile... 0 0 0.5 semitones re 1 Hz
semitone1stquant= Get quantile... 0 0 0.25 semitones re 1 Hz
semitone3rdquant= Get quantile... 0 0 0.75 semitones re 1 Hz
semitoneIQR=semitone3rdquant-semitone1stquant
semitonerange=semitonemax-semitonemin

semitonemean=Get mean... 0 0 semitones re 1 Hz
semitonestnddev=Get standard deviation... 0 0 semitones

#-----
hlogmin=Get minimum... 0 0 "Hertz (logarithmic)" Parabolic
hlogmax=Get maximum... 0 0 "Hertz (logarithmic)" Parabolic

hlogmean= Get mean... 0 0 Hertz (logarithmic)
hlogmedian= Get quantile... 0 0 0.5 Hertz (logarithmic)
hlog1stquant= Get quantile... 0 0 0.25 Hertz (logarithmic)
hlog3rdquant= Get quantile... 0 0 0.75 Hertz (logarithmic)
hlogIQR=hlog3rdquant-hlog1stquant
hlogrange=hlogmax-hlogmin

#Slope values in Hertz and Semitones

slope=Get mean absolute slope... Hertz
semitoneslope=Get mean absolute slope... Semitones
semitoneslopenooctave=Get slope without octave jumps


#-------------------------------------------------------------------------
-------
#this carries out the Intensity analysis with the minimum range value set to
50

select mySound
rms = Get root-mean-square... 0 0
select mySound
To Intensity... 50 0 yes
Select outer viewport... 0 12 0 6
Blue
Draw... 0 0 0 0 yes
Marks left every... 1 2 yes yes yes
Marks bottom every... 1 0.1 yes yes yes
Draw inner box

#This rites the graphic to a pdf file
Write to PDF file... Analysis_Files For Asset
'object_name$'/Intensity_Contour_File 'object_name$'.pdf
Erase all

intensitymin=Get minimum... 0 0 Parabolic
intensitymax=Get maximum... 0 0 Parabolic
intensitymean=Get mean... 0 0 energy
intensitystddev=Get standard deviation... 0 0
intensitymedian=Get quantile... 0 0 0.5
intensity1stquant=Get quantile... 0 0 0.25
intensity3rdquant=Get quantile... 0 0 0.75
intensityrange=intensitymax-intensitymin
intensityiqr=intensity3rdquant-intensity1stquant

#-------------------------------------------------------------------------
-------
#This gives a measure of spectral energy

select mySound
To Spectrum... yes
Select outer viewport... 0 12 0 6
Red
Erase all
Draw... 0 0 0 0 yes
```

```
Marks left every... 1 5 yes yes yes
Marks bottom every... 1 1000 yes yes yes
Draw inner box

Write to PDF file... Analysis_Files For Asset  'object_name$'/Sound_Spectrum
'object_name$'.pdf

spectralgravitycentre=Get centre of gravity... 2

spectralstnddev=Get standard deviation... 2

#-------------------------------------------------------------------------
-------
#This gets the point process for jitter and shimmer measurements
select mySound
To PointProcess (periodic, cc)... 75 600
Get jitter (local)... 0 0 0.0001 0.02 1.3

select mySound
plus PointProcess 'object_name$'
Get shimmer (local)... 0 0 0.0001 0.02 1.3 1.6

select mySound
plus PointProcess 'object_name$'
plus Pitch 'object_name$'
Voice report... 0 0 75 600 1.3 1.6 0.03 0.45

filedelete Analysis_Files For Asset  'object_name$'/ Voice Report For
'object_name$'.txt
fappendinfo Analysis_Files For Asset  'object_name$'/ Voice Report For
'object_name$'.txt


#-------------------------------------------------------------------------
-------

#-------------------------------------------------------------------------
-------
# This gets the Long Term Spectral Energy measurement (LTAS)
select mySound
To Ltas... 500

Select outer viewport... 0 12 0 6
Erase all
Draw... 0 0 -20 80 yes Bars
Red
Marks left every... 1 5 yes yes yes
Marks bottom every... 1 1000 yes yes yes

Write to PDF file... Analysis_Files For Asset
'object_name$'/Sound_LTAS_'object_name$'.pdf

#-------------------------------------------------------------------------
-------
#This creates a spectrogram and saves it to file

select mySound
To Spectrogram... 0.005 5000 0.002 20 Gaussian

Select outer viewport... 0 12 0 6
Erase all
Paint... 0 0 0 0 100 yes 50 6 0 yes
Marks left every...  1 500 yes yes yes
Marks bottom every...  1 0.1 yes yes yes
Draw inner box

Write to PDF file... Analysis_Files For Asset
'object_name$'/Spectrogram_'object_name$'.pdf
```

```
#------------------------------------------------------------------------
-------
#This writes another spectrogram but with the pitch and intensity contours
drawn on it also.

#THSI IS POINTLESS AS THE SCALES FOR THE DIFFERENT MEASUREMENTS DO NOT MATCH
# I have left it in but commented it out

#select mySound
#To Spectrogram... 0.005 5000 0.002 20 Gaussian

#Select outer viewport... 0 12 0 6
#Erase all
#Paint... 0 0 0 0 100 yes 50 6 0 yes
#Marks left every...  1 500 yes yes yes
#Marks bottom every...  1 0.1 yes yes yes
#Draw inner box

#Write to PDF file... Analysis_Files For Asset
'object_name$'/Spectrogram_'object_name$'.pdf

#Red
#select Pitch 'object_name$'
#Draw... 0 0 0 q3*1.5 yes
#Blue
#select Intensity 'object_name$'
#Draw... 0 0 0 0 yes


#Write to PDF file... Analysis_Files For Asset
'object_name$'/Spectrogram_Intensity)Pitch_Contours'object_name$'.pdf

#------------------------------------------------------------------------
-------
#This writes the pitch values to the text file

clearinfo
printline Pitch Info for 'object_name$':
printline _____________________________________________
printline Minimum  Pitch  'tab$' 'tab$' Maximum Pitch  'tab$'   Pitch Range
printline 'pitchmin'   'tab$' 'pitchmax'   'tab$' 'tab$' 'pitchrange'
printline
printline Pitch Mean  'tab$' 'tab$'  'tab$' Pitch Median
printline 'pitchmean' Hz   'tab$' 'pitchmedian' Hz
printline
printline 25th Percentile  'tab$'  'tab$' 75th Percentile
printline 'pitch1stquant' Hz  'tab$'   'pitch3rdquant' Hz
printline
printline Pitch Stnd. Dev.  'tab$'   Pitch IQR
printline 'pitchstnddev'  Hz 'tab$'   'pitchiqr' Hz
printline
printline  _____________________________________________
printline Semitone values
printline
printline  Semitone min 'tab$' Semitone max 'tab$'        Semitone Range
printline  'semitonemin' st 'tab$' 'semitonemax' 'tab$' 'semitonerange'
printline
printline Semitone Mean 'tab$' Semitone Median
printline 'semitonemean' st 'tab$'  'semitonemedian'
printline
printline Semitone 25th percentile 'tab$' Semitone 75th percentile
printline 'semitone1stquant' 'tab$'    'semitone3rdquant'
printline
printline Semitone Stand. Dev 'tab$' Semitone IQR
printline 'semitonestnddev' 'tab$' 'semitoneIQR'
printline
printline _____________________________________________
```

```
printline Slope information
printline
printline Mean Slope in Hertz
printline  'slope' Hz/s
printline Mean Slope in Semitones
printline
printline 'semitoneslope' semitones/s
printline
printline Semitone slope without Octave jumps
printline 'semitoneslopenooctave' semitones/s
printline _____________________________________________
printline
printline Hertz(Logarithmic) Values: equivalent to calculating semitone
values and converting back to Hertz
printline
printline Hertz(log) min 'tab$' Hertz(log) max 'tab$'        Hertz(log)
Range
printline 'hlogmin' 'tab$' 'hlogmax' 'tab$'    'hlogrange'
printline
printline Hert(log) Mean 'tab$' Hertz(log) Median
printline 'hlogmean' 'tab$' 'hlogmedian'
printline
printline Hertz(log) 25th percentile 'tab$' Hertz(log) 75th Percentile
printline 'hlog1stquant' 'tab$'        'hlog3rdquant'
printline
printline Hertz(log) IQR
printline 'hlogIQR'
printline
printline _____________________________________________

#This writes Intensity values to an info window to be saved as a file.

printline
printline Intensity Info 'object_name$':
printline _____________________________________________
printline Minimum  Intensity  'tab$' 'tab$' Maximum Intensity  'tab$'
Intensity Range
printline 'intensitymin' dB 'tab$' 'intensitymax' dB 'tab$'
'intensityrange' dB
printline
printline Intensity Mean   'tab$' 'tab$'  'tab$' Intensity Median
printline 'intensitymean' dB 'tab$' 'intensitymedian' dB
printline
printline 25th Percentile  'tab$' 'tab$' 75th Percentile
printline 'intensity1stquant' dB 'tab$' 'intensity3rdquant' dB
printline
printline Intensity Stnd. Dev  'tab$' 'tab$'  Intensity IQR
printline 'intensitystddev' dB 'tab$' 'intensityiqr' dB
printline _____________________________________________
printline
printline Root Mean Square Energy
printline 'rms'
printline
printline _____________________________________________


#This writes the spectral centre of gravity to the file

printline
printline Spectral energy Info for 'object_name$':
printline _____________________________________________
printline  Centre of Gravity  'tab$'  Spectral Standard Deviation
printline 'spectralgravitycentre' Hz 'tab$' 'spectralstnddev' hZ
printline _____________________________________________

#This deselects the sound object. Probably not necessary but
#we are done with it so its just to be safe to avoid errors.
minus Sound 'object_name$'
```

```
#This deletes the file in the subfolder so a new one can be written.
filedelete Analysis_Files For Asset  'object_name$'/
Acoustic_Info_for_Asset'object_name$'.txt



fappendinfo Analysis_Files For Asset  'object_name$'/
Acoustic_Info_for_Asset'object_name$'.txt


select mySound
Write to AIFF file... Analysis_Files For Asset
'object_name$'/'object_name$'.aiff


i=i+1

endwhile
```

Appendix J. PRAAT Script Used For The Extraction Of Syllable Nuclei From Speech Assets For Calculating Speech Rate

```
#############################################################################
#                                                                           #
#  Praat Script Syllable Nuclei                                             #
#  Copyright (C) 2008  Nivja de Jong and Ton Wempe                          #
#                                                                           #
#    This program is free software: you can redistribute it and/or modify   #
#    it under the terms of the GNU General Public License as published by   #
#    the Free Software Foundation, either version 3 of the License, or      #
#    (at your option) any later version.                                    #
#                                                                           #
#    This program is distributed in the hope that it will be useful,        #
#    but WITHOUT ANY WARRANTY; without even the implied warranty of         #
#    MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the          #
#    GNU General Public License for more details.                           #
#                                                                           #
#    You should have received a copy of the GNU General Public License      #
#    along with this program.  If not, see http://www.gnu.org/licenses/     #
#                                                                           #
#############################################################################


# counts syllables of all sound utterances in a directory
# NB unstressed syllables are sometimes overlooked
# NB filter sounds that are quite noisy beforehand
# NB use Ignorance Level/Intensity Median (dB) = 0 for unfiltered sounds,
#    use 2 for filtered sounds
# NB use Minimum dip between peaks (dB) = 2 for unfiltered sounds,
#    use 4 for filtered sounds


form Counting Syllables in Sound Utterances
   real Ignorance_Level/Intensity_Median_(dB) 0 or 2
   positive Max_number_of_syllables_per_Sound 700
   real Minimum_dip_between_peaks_(dB) 2 or 4
   boolean display_name yes
   sentence directory C:\directorywithsoundfiles
endform



# shorten variables
iglevel = 'ignorance_Level/Intensity_Median'
maxsyl = 'max_number_of_syllables_per_Sound'
mindip = 'minimum_dip_between_peaks'



Create Strings as file list... list 'directory$'\*.wav
numberOfFiles = Get number of strings
for ifile to numberOfFiles
   select Strings list
```

```
fileName$ = Get string... ifile
Read from file... 'directory$'\'fileName$'


sr = Get sample rate
# stay some distance from nyquist
bw = sr * 0.45


obj$ = selected$("Sound")
originaldur = Get total duration


Subtract mean


# Use intensity to get threshold
To Intensity... 50 0 yes
start = Get time from frame number... 1
nframes = Get number of frames
end = Get time from frame number... 'nframes'


# estimate noise floor
minint = Get minimum... 0 0 Parabolic


# estimate noise max
maxint = Get maximum... 0 0 Parabolic


#get median of Intensity: limits influence of high peaks
medint = Get quantile... 0 0 0.5
# estimate Intensity threshold
threshold = medint + iglevel
if threshold < minint
    threshold = minint
endif


Down to Matrix
# Convert intensity to sound
To Sound (slice)... 1
Rename... int


intdur = Get finishing time
intmax = Get maximum... 0 0 Parabolic


# estimate peak positions (all peaks)
To PointProcess (extrema)... Left yes no Sinc70


numpeaks = Get number of points


# fill array with time points
for i from 1 to numpeaks
    t'i' = Get time from index... 'i'
endfor
```

```
# fill array with intensity values
select Sound int
peakcount = 0
for i from 1 to numpeaks
    value = Get value at time... t'i' Cubic
    if value > threshold
          peakcount += 1
          int'peakcount' = value
          timepeaks'peakcount' = t'i'
    endif
endfor


# fill array with valid peaks: only intensity values if preceding
# dip in intensity is greater than mindip
select Intensity 'obj$'
validpeakcount = 0
precedingtime = timepeaks1
precedingint = int1
for p to peakcount-1
   following = p + 1
   followingtime = timepeaks'following'
   dip = Get minimum... 'precedingtime' 'followingtime' None
   diffint = abs(precedingint - dip)
   if diffint > mindip
      validpeakcount += 1
      validtime'validpeakcount' = timepeaks'p'
      precedingtime = timepeaks'following'
      precedingint = Get value at time... timepeaks'following' Cubic
   endif
endfor


# Look for only voiced parts
select Sound 'obj$'
To Pitch (ac)... 0.02 30 4 no 0.03 0.25 0.01 0.35 0.25 450


voicedcount = 0
for i from 1 to validpeakcount
   querytime = validtime'i'


   value = Get value at time... 'querytime' Hertz Linear


   if value <> undefined
      voicedcount = voicedcount + 1
      voicedpeak'voicedcount' = validtime'i'
   endif
endfor


# stop if too many syllables are found
if voicedcount > maxsyl
    pause 'obj$': Number of syllabes exceeds 'maxsyl'!
    exit
endif

# calculate time correction due to shift in time for Sound object versus
# intensity object
timecorrection = originaldur/intdur
```

```
    # Insert voiced peaks in second Tier
    select Sound 'obj$'
    To TextGrid... "syllables" syllables
    for i from 1 to voicedcount
        position = voicedpeak'i' * timecorrection
        Insert point... 1 position 'i'
    endfor


    # write textgrid to textfile
    Write to text file... 'directory$'\'obj$'.syllables.TextGrid


    # clean up before next sound file is opened
    select all
    minus Strings list
    Remove


endfor
```

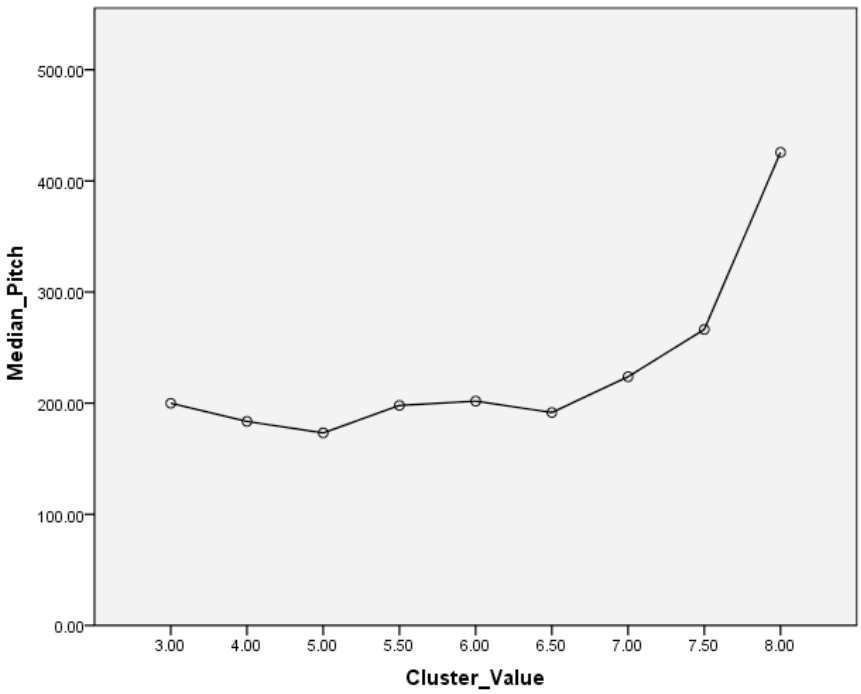
Appendix K. Reports For The Spearman's Rank Correlation Procedure For The Activation Dimension With Trend Line Scatter Plots

Median pitch

Correlations

			Cluster_Value	Median_Pitch
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.750*
		Sig. (2-tailed)	.	.020
		N	9	9
	Median_Pitch	Correlation Coefficient	.750*	1.000
		Sig. (2-tailed)	.020	.
		N	9	9

*. Correlation is significant at the 0.05 level (2-tailed).

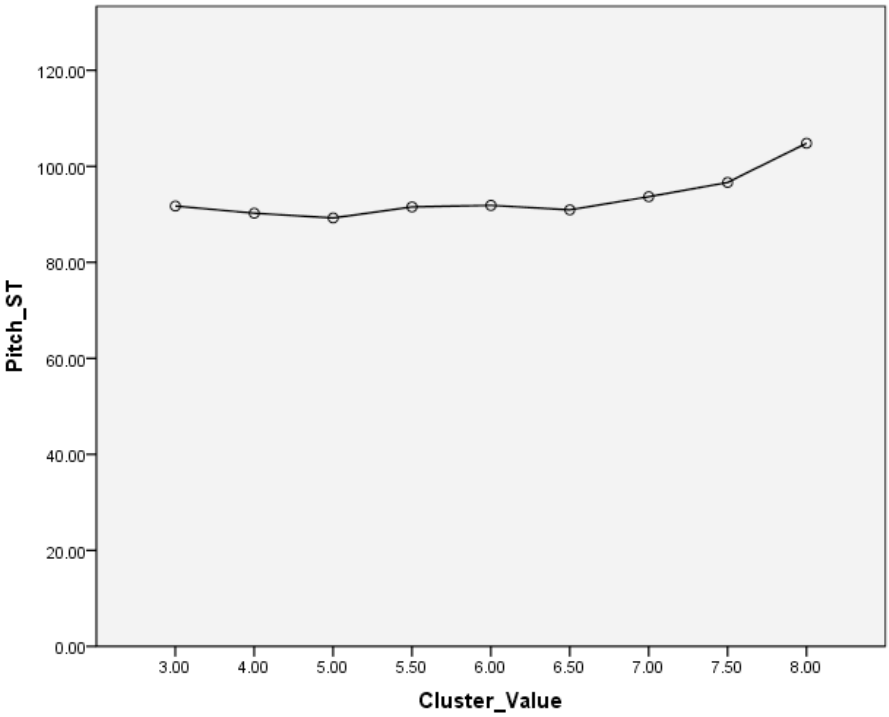


Median pitch in semitones

Correlations

			Cluster_Value	Pitch_ST
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.750*
		Sig. (2-tailed)	.	.020
		N	9	9
	Pitch_ST	Correlation Coefficient	.750*	1.000
		Sig. (2-tailed)	.020	.
		N	9	9

*. Correlation is significant at the 0.05 level (2-tailed).

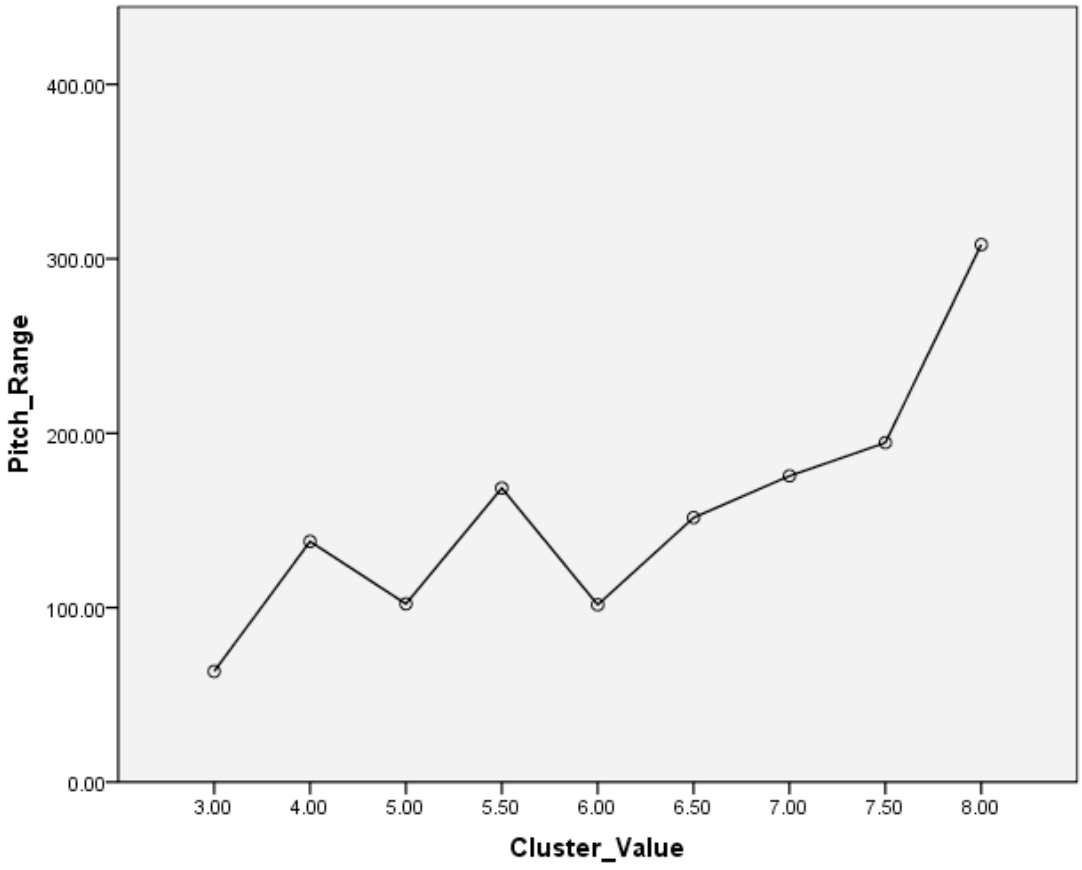


Pitch range

Correlations

			Cluster_Value	Pitch_Range
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.850**
		Sig. (2-tailed)	.	.004
		N	9	9
	Pitch_Range	Correlation Coefficient	.850**	1.000
		Sig. (2-tailed)	.004	.
		N	9	9

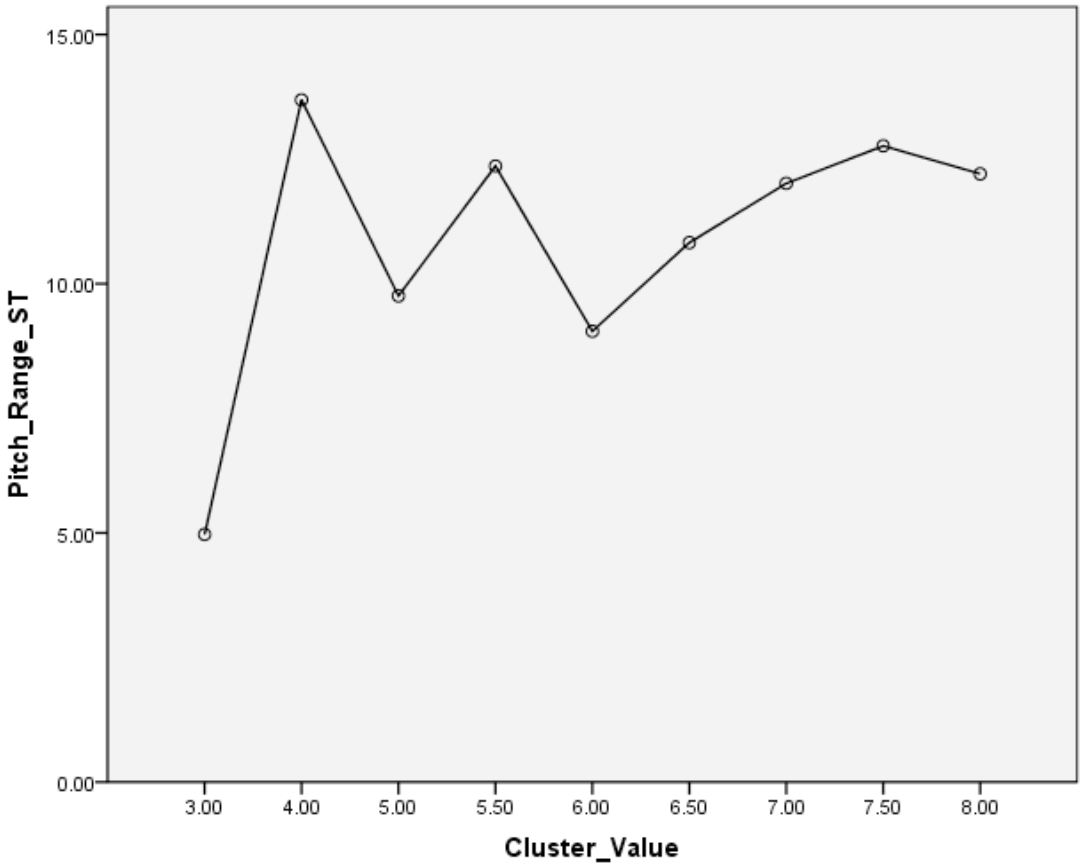
**. Correlation is significant at the 0.01 level (2-tailed).



Pitch range in semitones

Correlations

			Cluster_Value	Pitch_Range_ST
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.300
		Sig. (2-tailed)	.	.433
		N	9	9
	Pitch_Range_ST	Correlation Coefficient	.300	1.000
		Sig. (2-tailed)	.433	.
		N	9	9

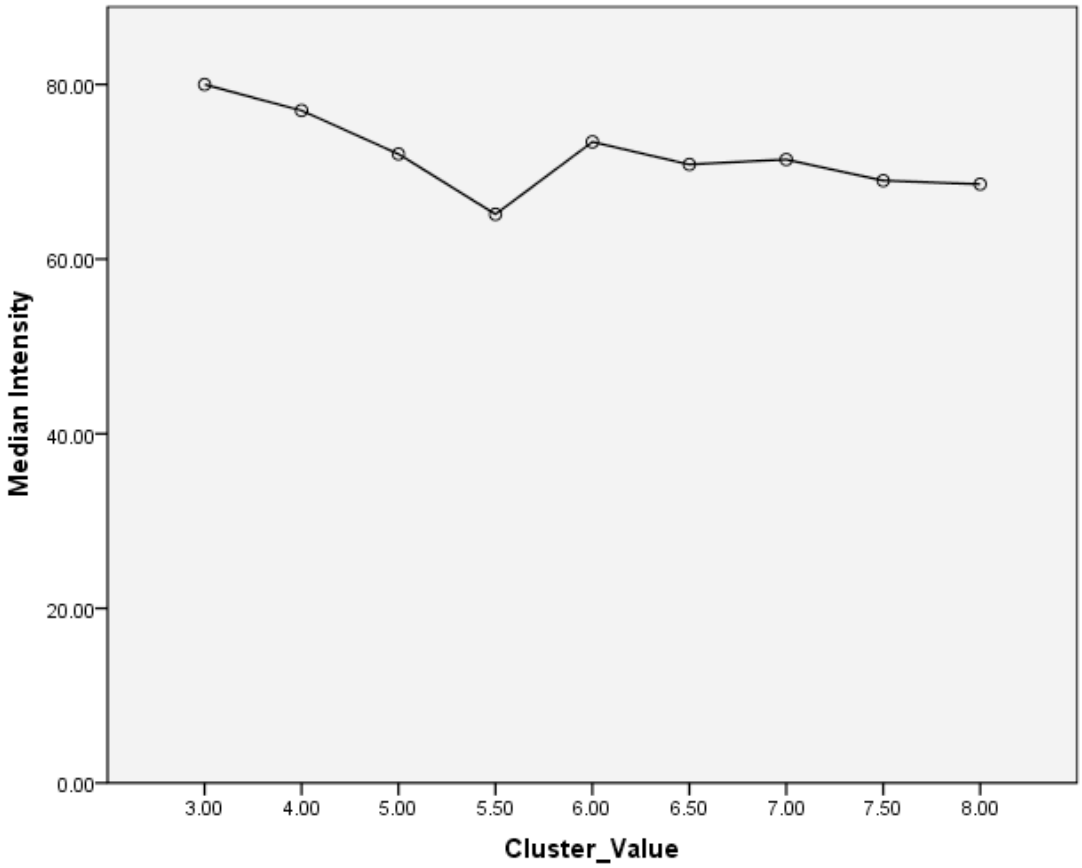


Median intensity

Correlations

			Cluster_Value	Intensity
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	-.700*
		Sig. (2-tailed)	.	.036
		N	9	9
	Intensity	Correlation Coefficient	-.700*	1.000
		Sig. (2-tailed)	.036	.
		N	9	9

*. Correlation is significant at the 0.05 level (2-tailed).

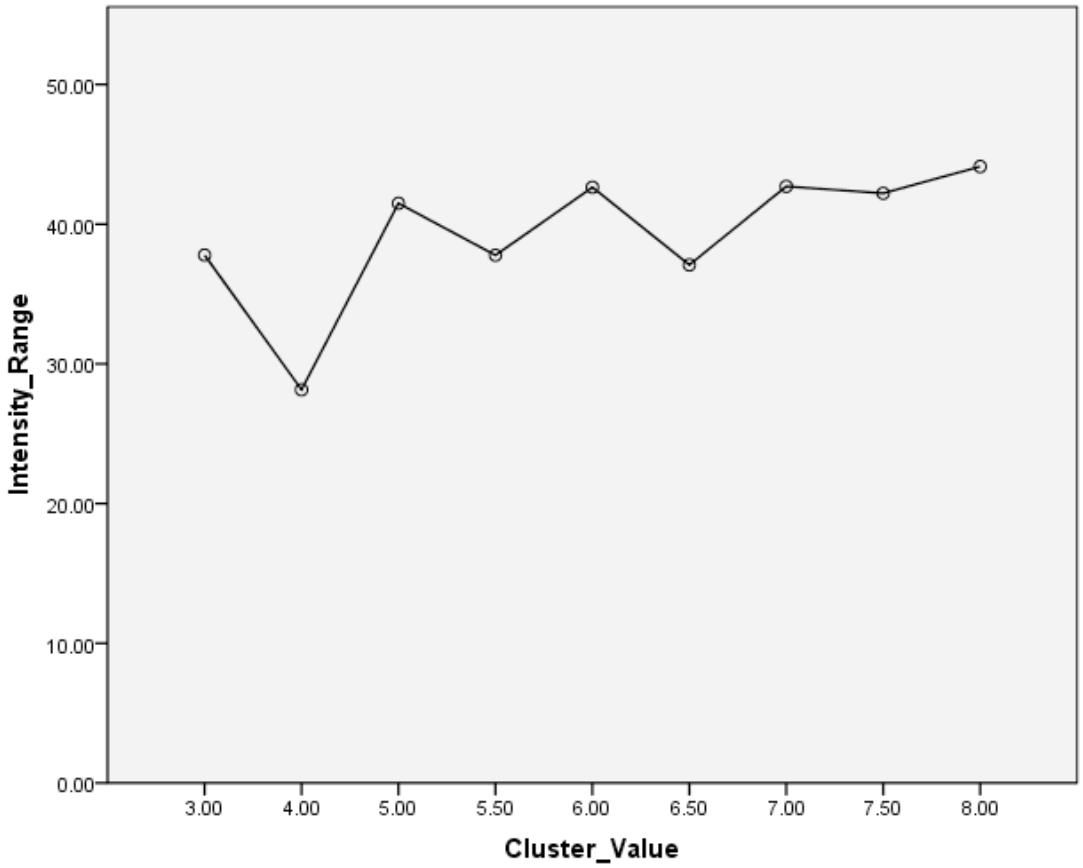


Intensity range

Correlations

			Cluster_Value	Intensity_Range
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.667*
		Sig. (2-tailed)	.	.050
		N	9	9
	Intensity_Range	Correlation Coefficient	.667*	1.000
		Sig. (2-tailed)	.050	.
		N	9	9

*. Correlation is significant at the 0.05 level (2-tailed).

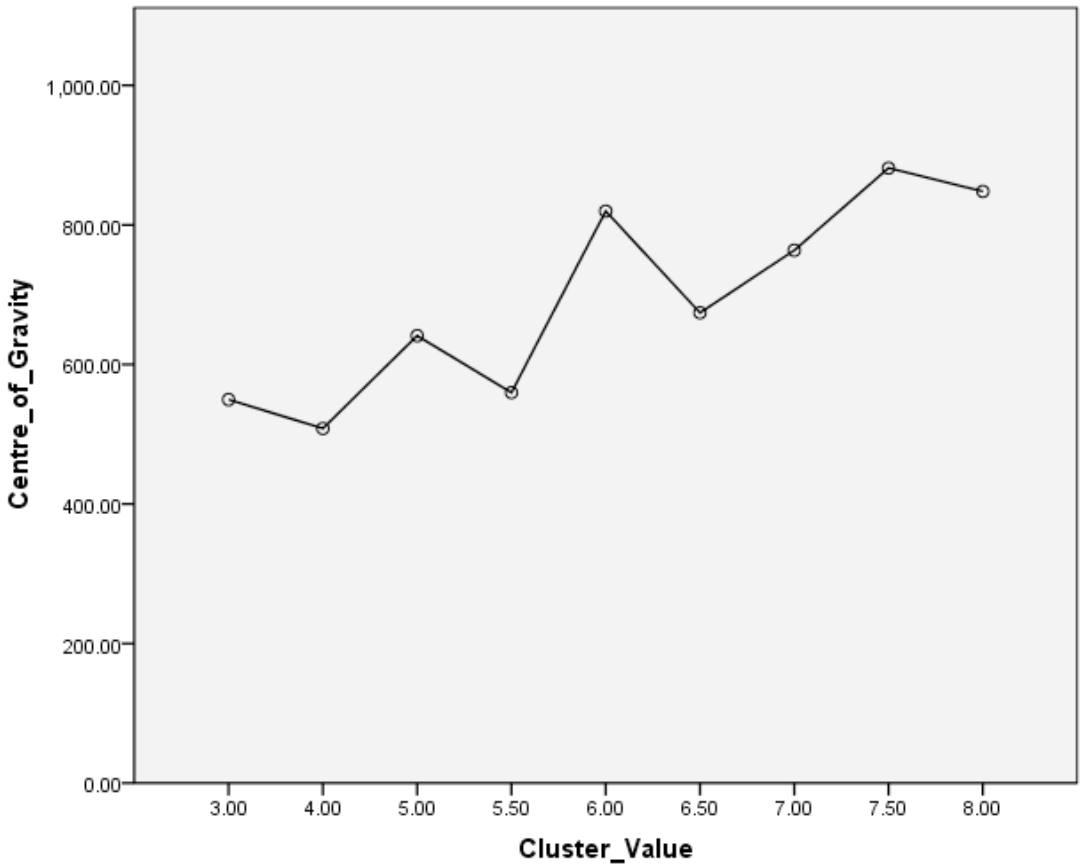


Centre of gravity

Correlations

			Cluster_Value	Centre_of_Gravity
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.900**
		Sig. (2-tailed)	.	.001
		N	9	9
	Centre_of_Gravity	Correlation Coefficient	.900**	1.000
		Sig. (2-tailed)	.001	.
		N	9	9

**. Correlation is significant at the 0.01 level (2-tailed).

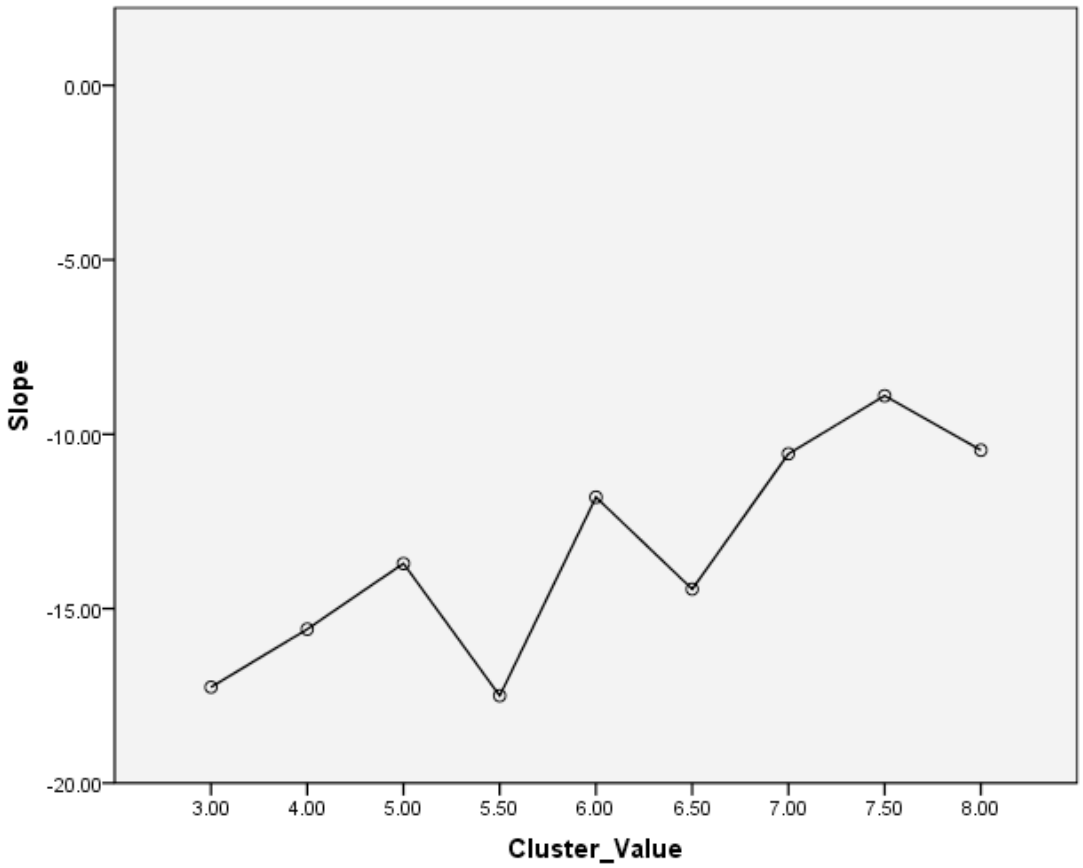


Spectral slope

Correlations

			Cluster_Value	Slope
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.817**
		Sig. (2-tailed)	.	.007
		N	9	9
	Slope	Correlation Coefficient	.817**	1.000
		Sig. (2-tailed)	.007	.
		N	9	9

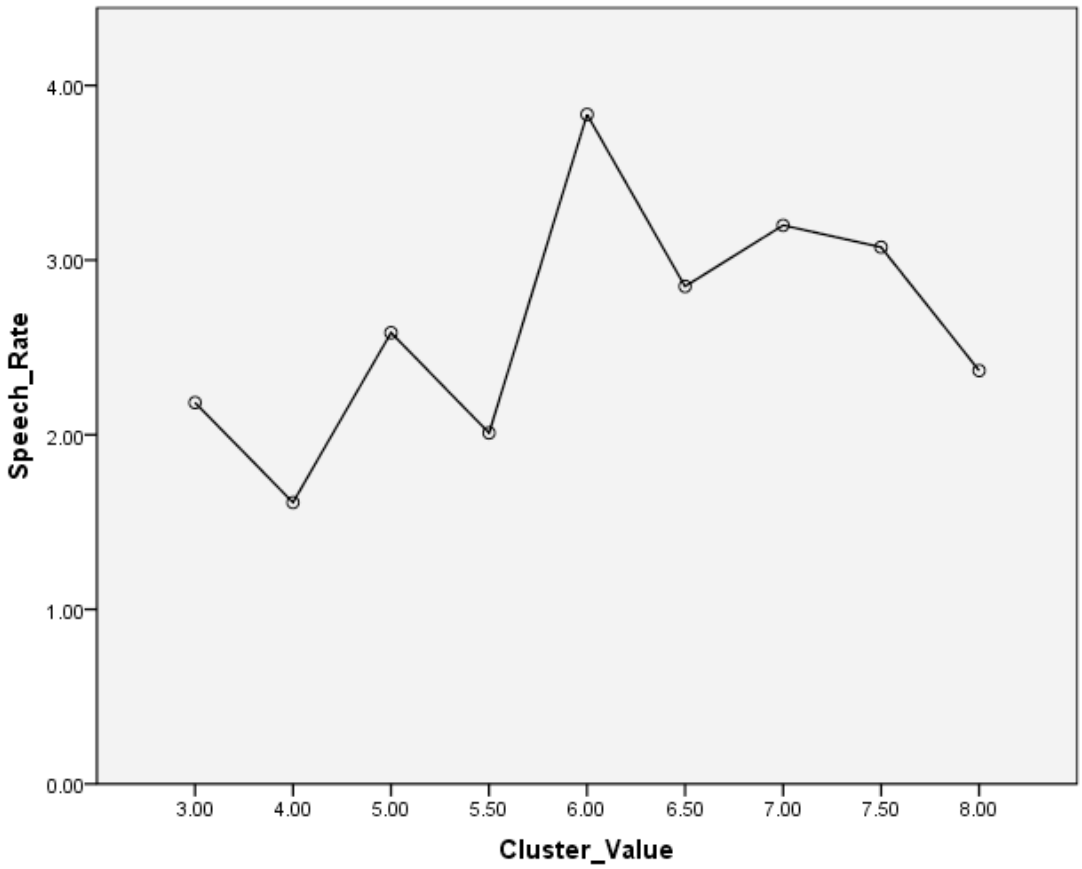
**. Correlation is significant at the 0.01 level (2-tailed).



Speech rate

Correlations

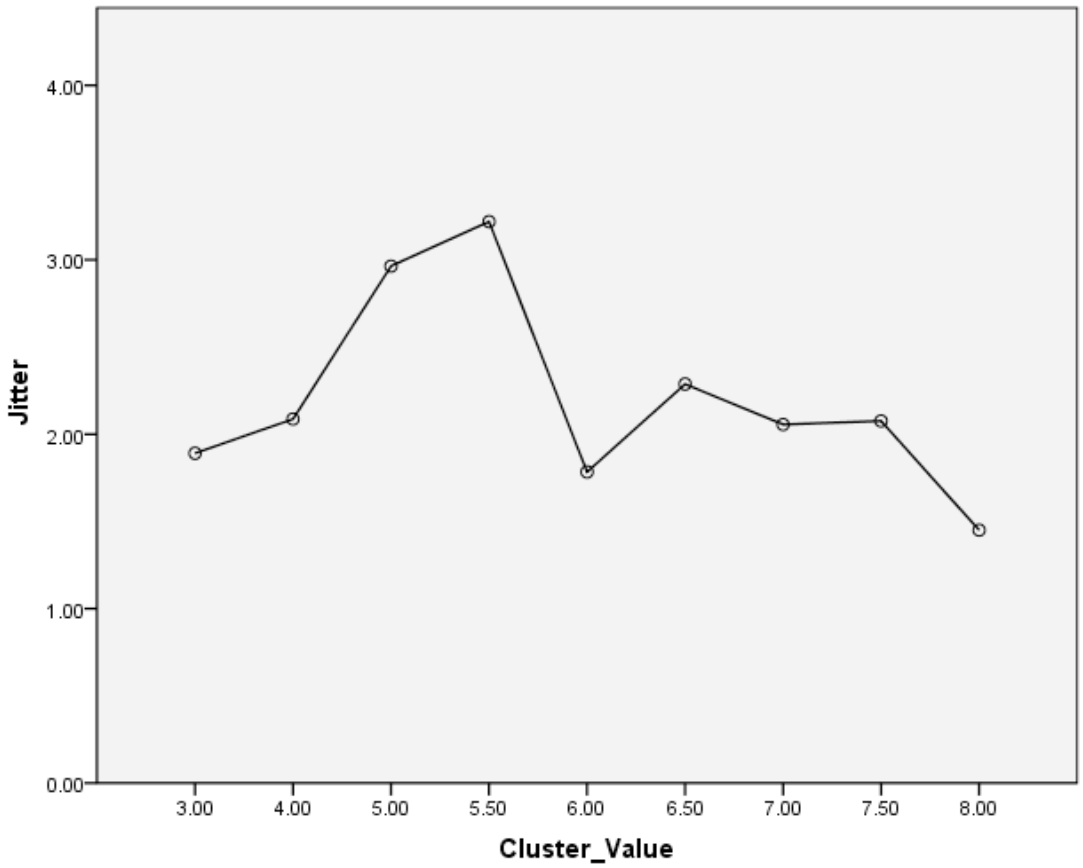
			Cluster_Value	Speech_Rate
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.533
		Sig. (2-tailed)	.	.139
		N	9	9
	Speech_Rate	Correlation Coefficient	.533	1.000
		Sig. (2-tailed)	.139	.
		N	9	9



Jitter

Correlations

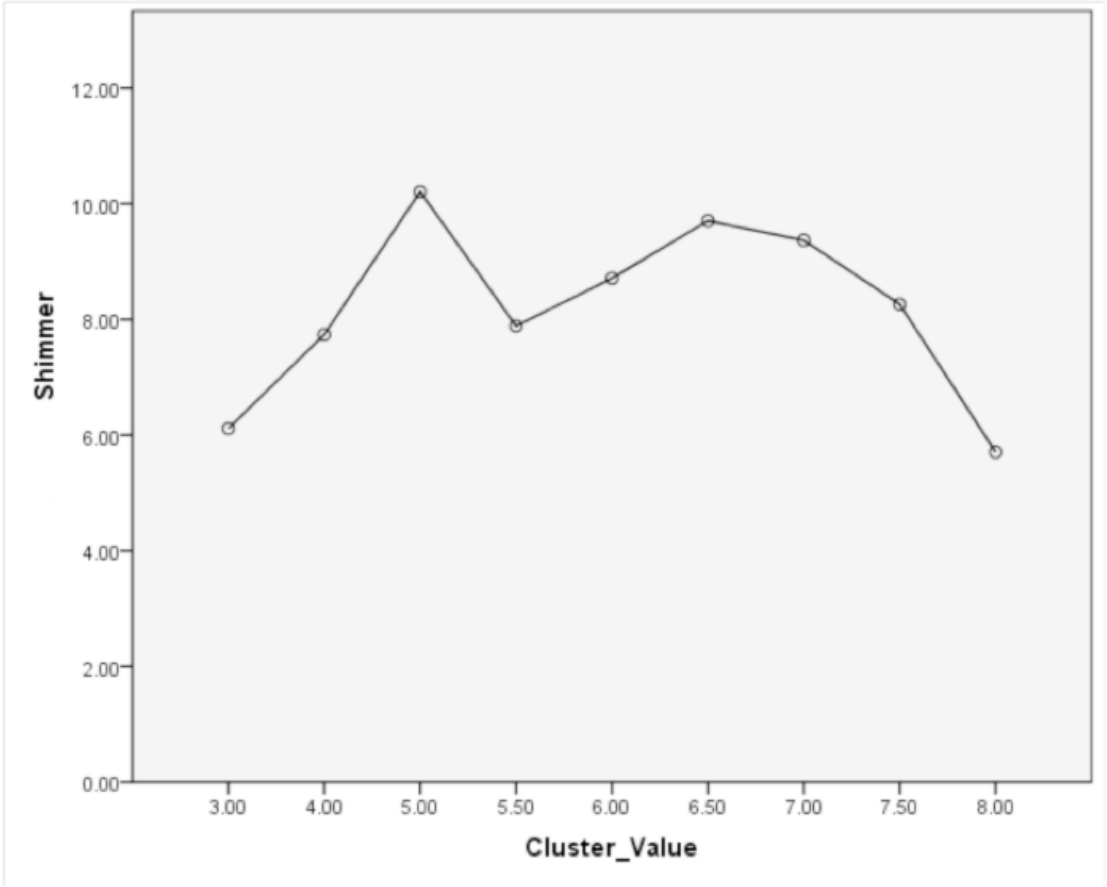
			Cluster_Value	Jitter
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	-.350
		Sig. (2-tailed)	.	.356
		N	9	9
	Jitter	Correlation Coefficient	-.350	1.000
		Sig. (2-tailed)	.356	.
		N	9	9



Shimmer

Correlations

			Cluster_Value	Shimmer
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.033
		Sig. (2-tailed)	.	.932
		N	9	9
	Shimmer	Correlation Coefficient	.033	1.000
		Sig. (2-tailed)	.932	.
		N	9	9



Intensity minimum and maximum

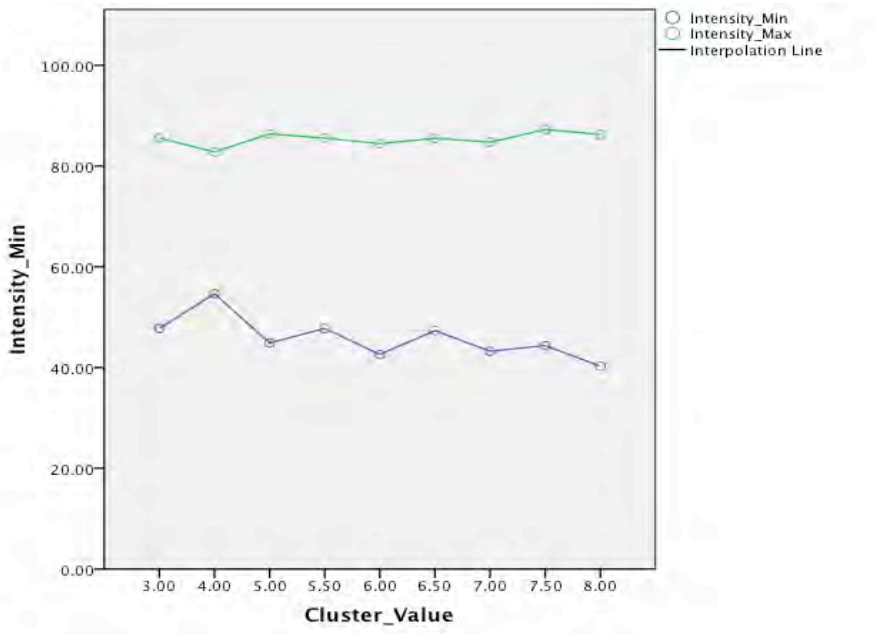
Correlations

			Cluster_Value	Intensity_Min
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	-.750*
		Sig. (2-tailed)	.	.020
		N	9	9
	Intensity_Min	Correlation Coefficient	-.750*	1.000
		Sig. (2-tailed)	.020	.
		N	9	9

*. Correlation is significant at the 0.05 level (2-tailed).

Correlations

			Cluster_Value	Intensity_Max
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.333
		Sig. (2-tailed)	.	.381
		N	9	9
	Intensity_Max	Correlation Coefficient	.333	1.000
		Sig. (2-tailed)	.381	.
		N	9	9

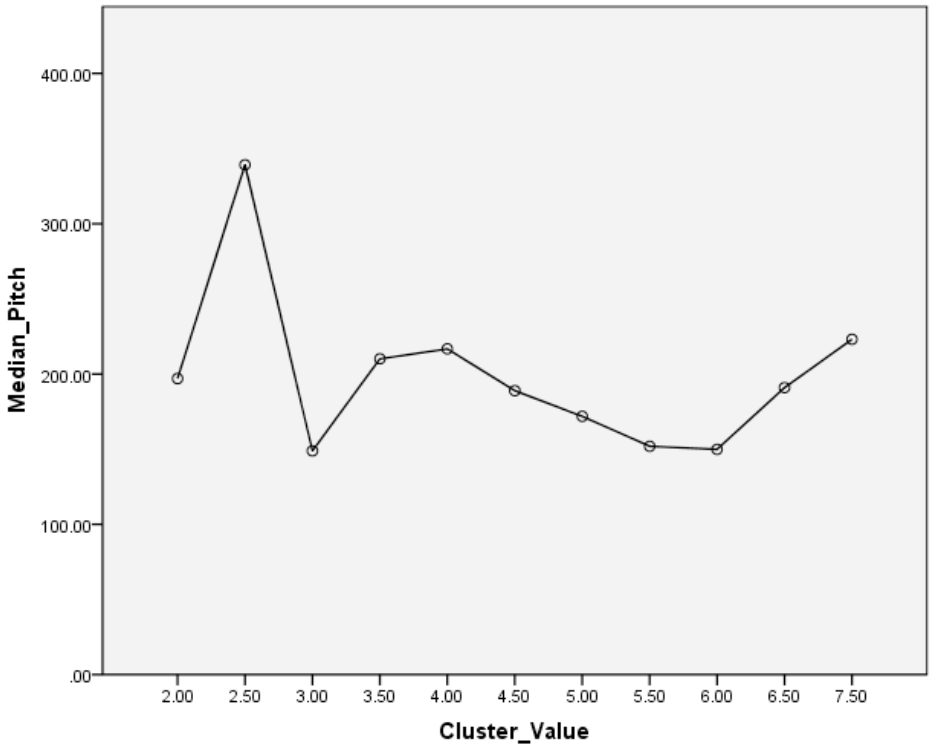


Appendix L. Reports For The Spearman's Rank Correlation Procedure For The Evaluation Dimension With Trend Line Scatter Plots

Median pitch

Correlations

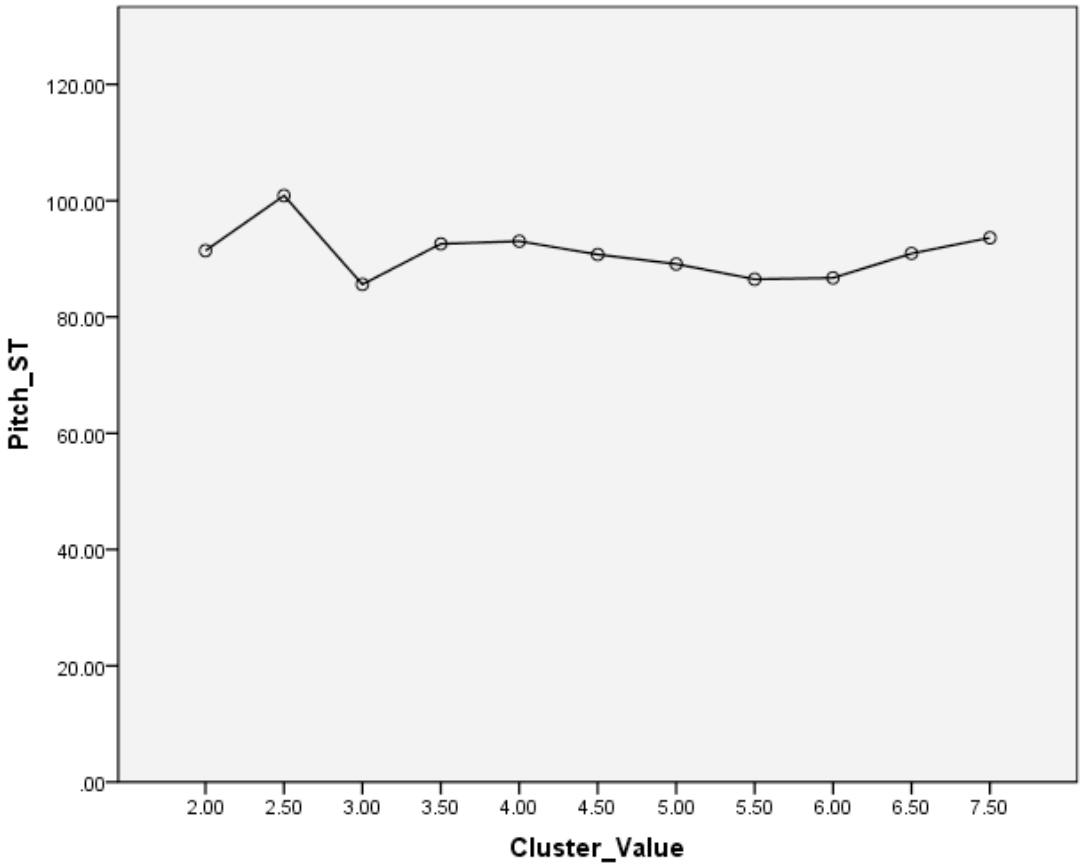
			Cluster_Value	Median_Pitch
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	-.155
		Sig. (2-tailed)	.	.650
		N	11	11
	Median_Pitch	Correlation Coefficient	-.155	1.000
		Sig. (2-tailed)	.650	.
		N	11	11



Median pitch in semitones

Correlations

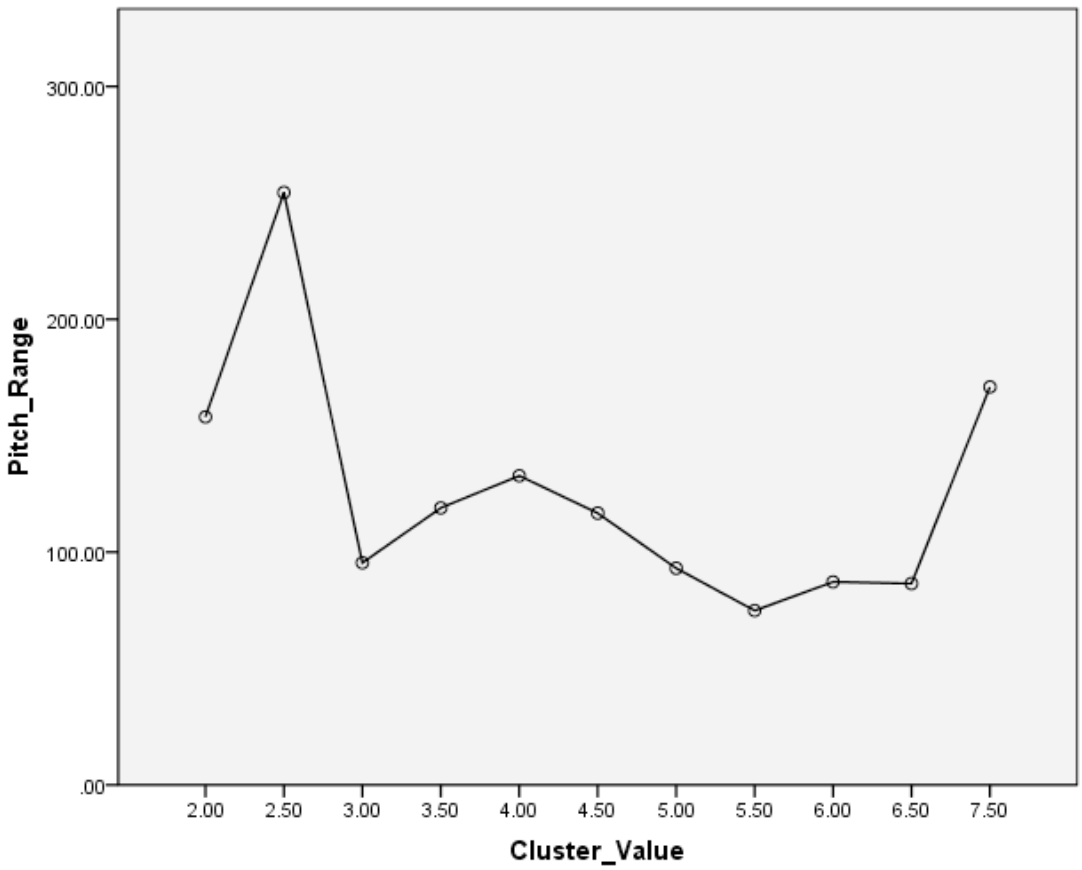
			Cluster_Value	Pitch_ST
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	-.145
		Sig. (2-tailed)	.	.670
		N	11	11
	Pitch_ST	Correlation Coefficient	-.145	1.000
		Sig. (2-tailed)	.670	.
		N	11	11



Pitch range

Correlations

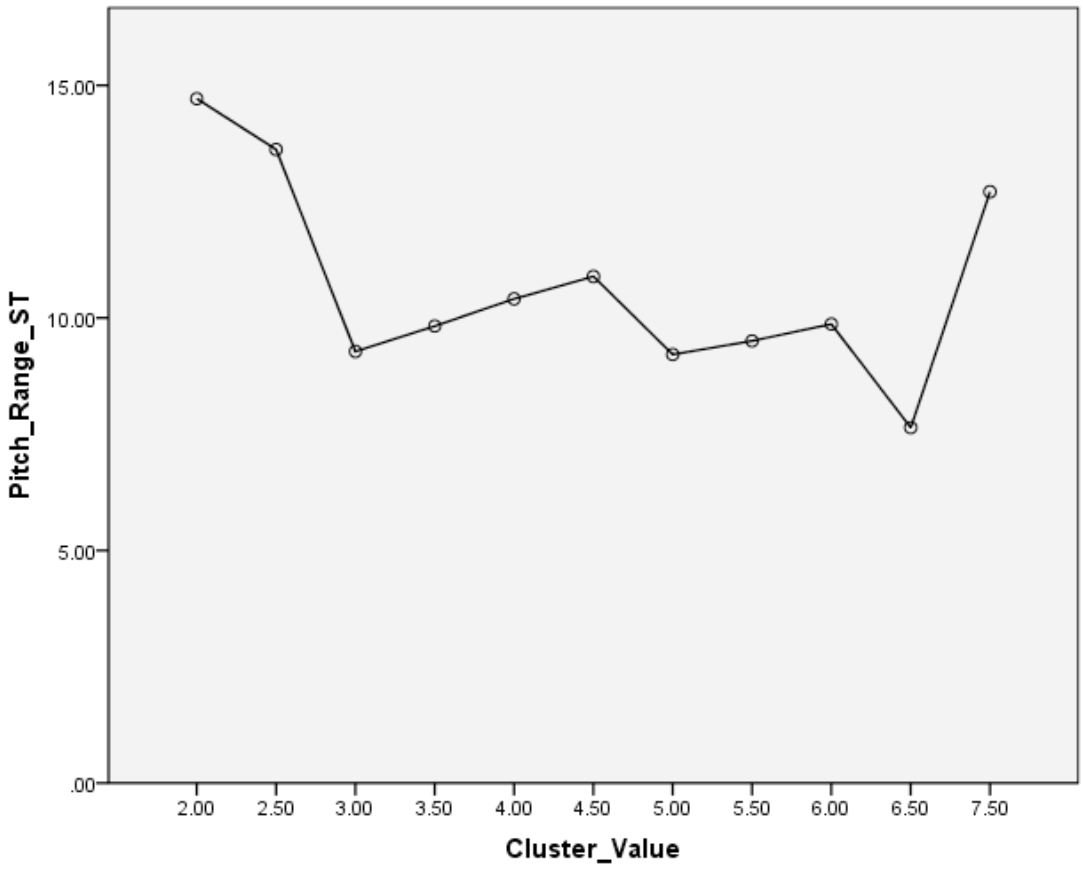
			Cluster_Value	Pitch_Range
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	-.482
		Sig. (2-tailed)	.	.133
		N	11	11
	Pitch_Range	Correlation Coefficient	-.482	1.000
		Sig. (2-tailed)	.133	.
		N	11	11



Pitch range in semitones

Correlations

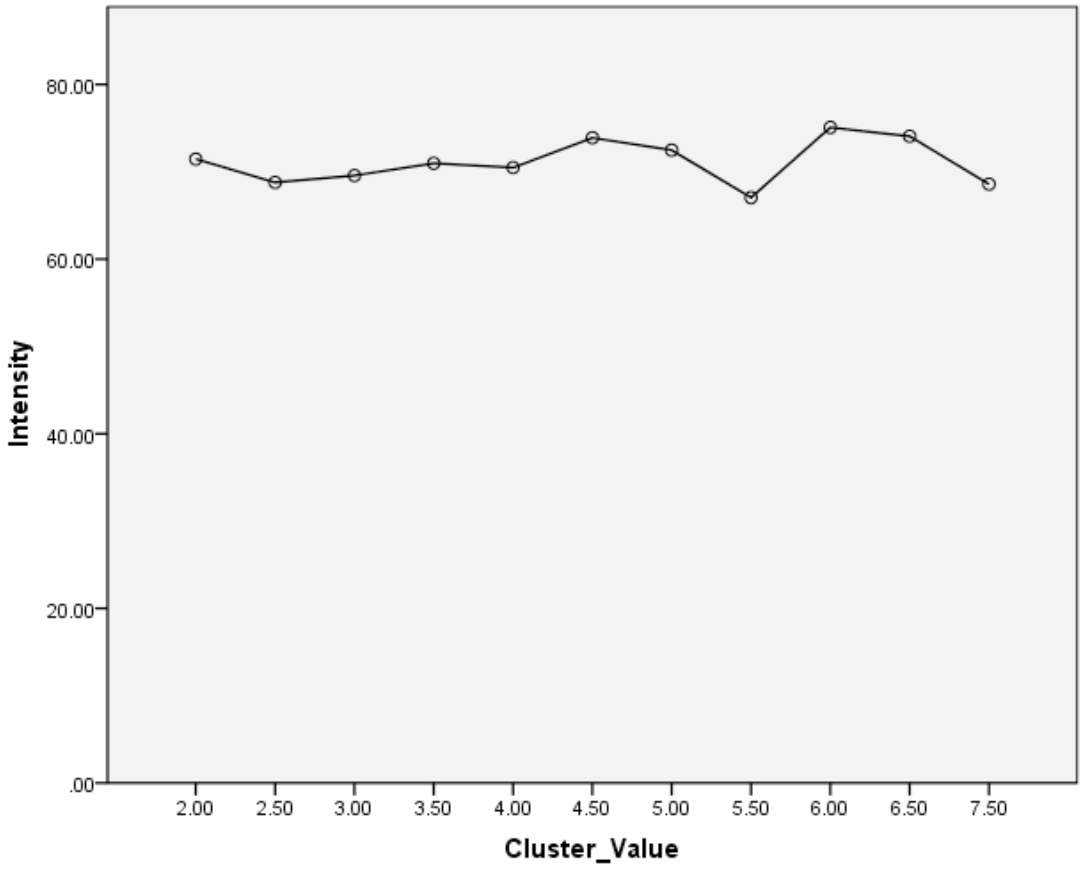
			Cluster_Value	Pitch_Range_ST
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	-.400
		Sig. (2-tailed)	.	.223
		N	11	11
	Pitch_Range_ST	Correlation Coefficient	-.400	1.000
		Sig. (2-tailed)	.223	.
		N	11	11



Intensity

Correlations

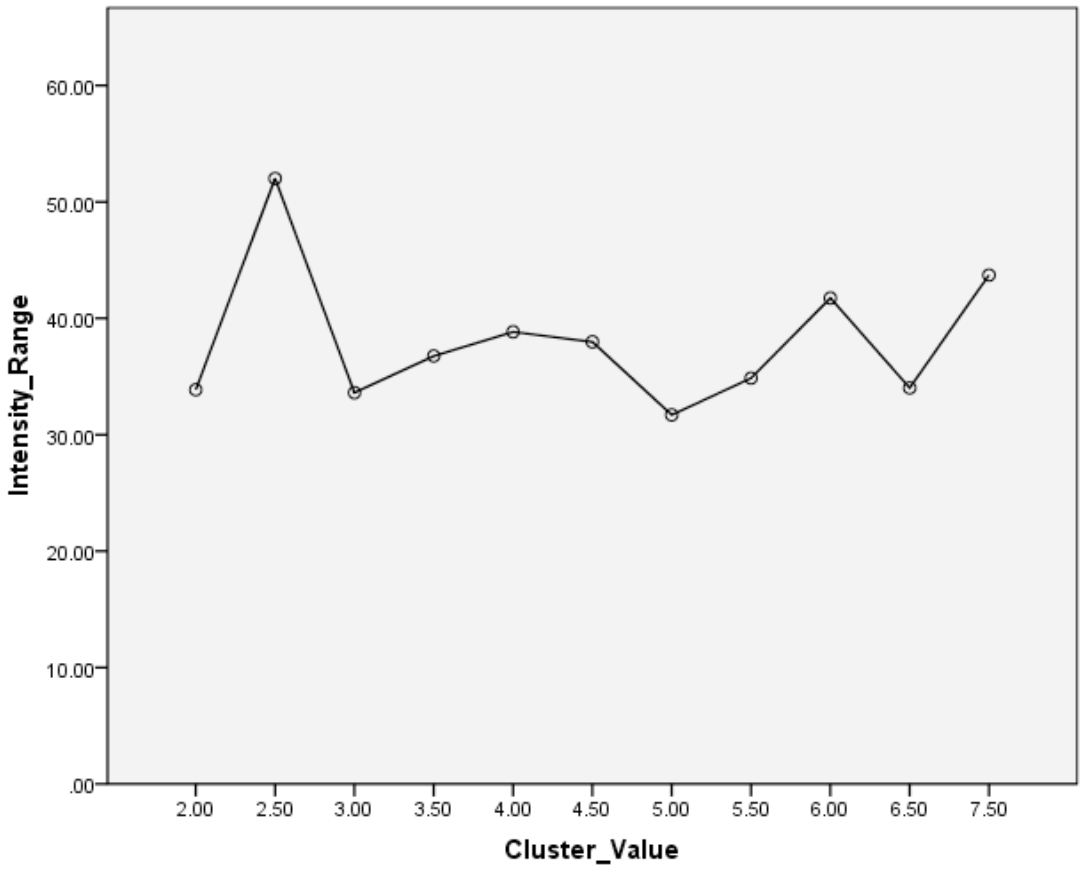
			Cluster_Value	Intensity
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.155
		Sig. (2-tailed)	.	.650
		N	11	11
	Intensity	Correlation Coefficient	.155	1.000
		Sig. (2-tailed)	.650	.
		N	11	11



Intensity range

Correlations

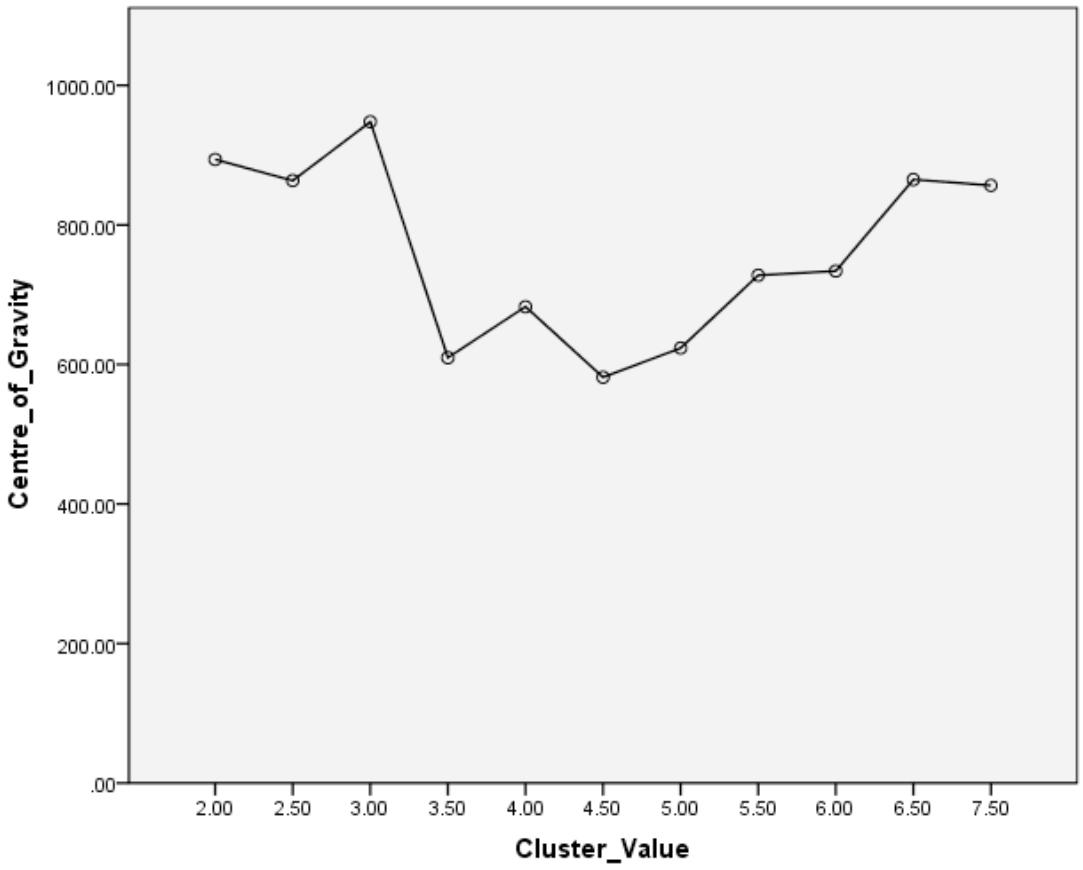
			Cluster_Value	Intensity_Range
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.173
		Sig. (2-tailed)	.	.612
		N	11	11
	Intensity_Range	Correlation Coefficient	.173	1.000
		Sig. (2-tailed)	.612	.
		N	11	11



Centre of gravity

Correlations

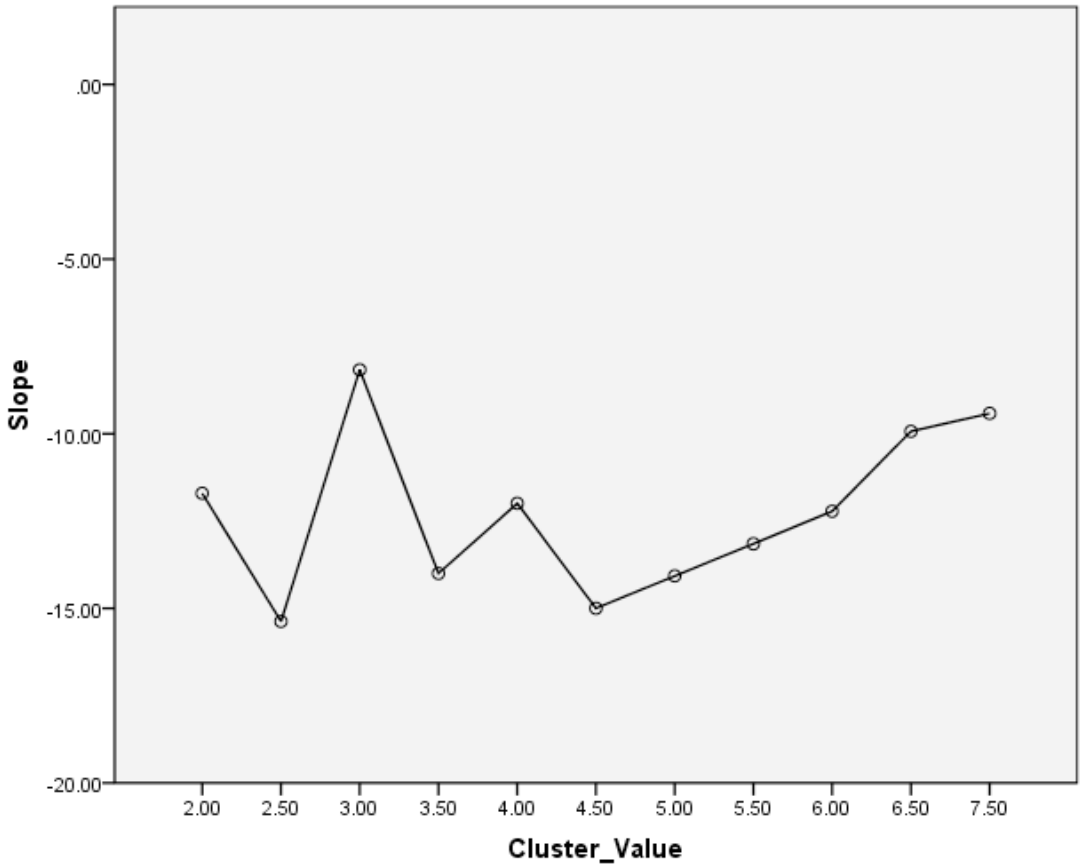
			Cluster_Value	Centre_of_Gravity
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	-.191
		Sig. (2-tailed)	.	.574
		N	11	11
	Centre_of_Gravity	Correlation Coefficient	-.191	1.000
		Sig. (2-tailed)	.574	.
		N	11	11



Spectral slope

Correlations

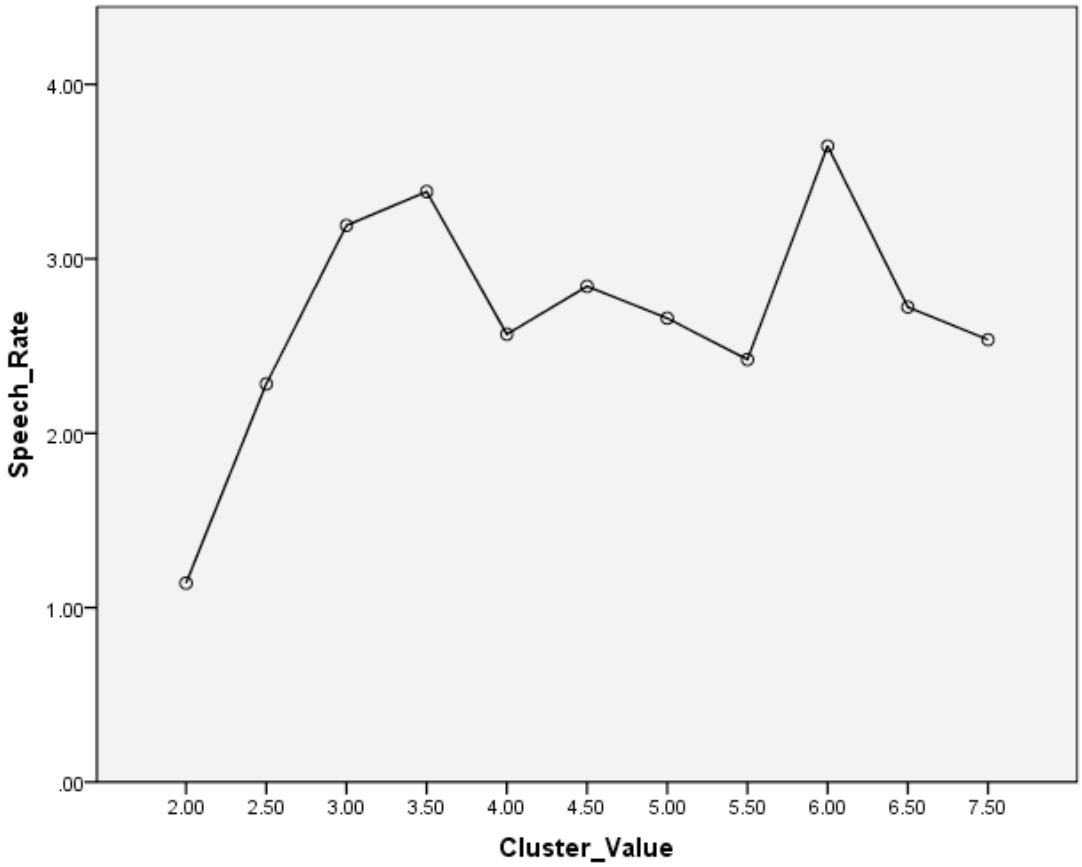
			Cluster_Value	Slope
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.227
		Sig. (2-tailed)	.	.502
		N	11	11
	Slope	Correlation Coefficient	.227	1.000
		Sig. (2-tailed)	.502	.
		N	11	11



Speech rate

Correlations

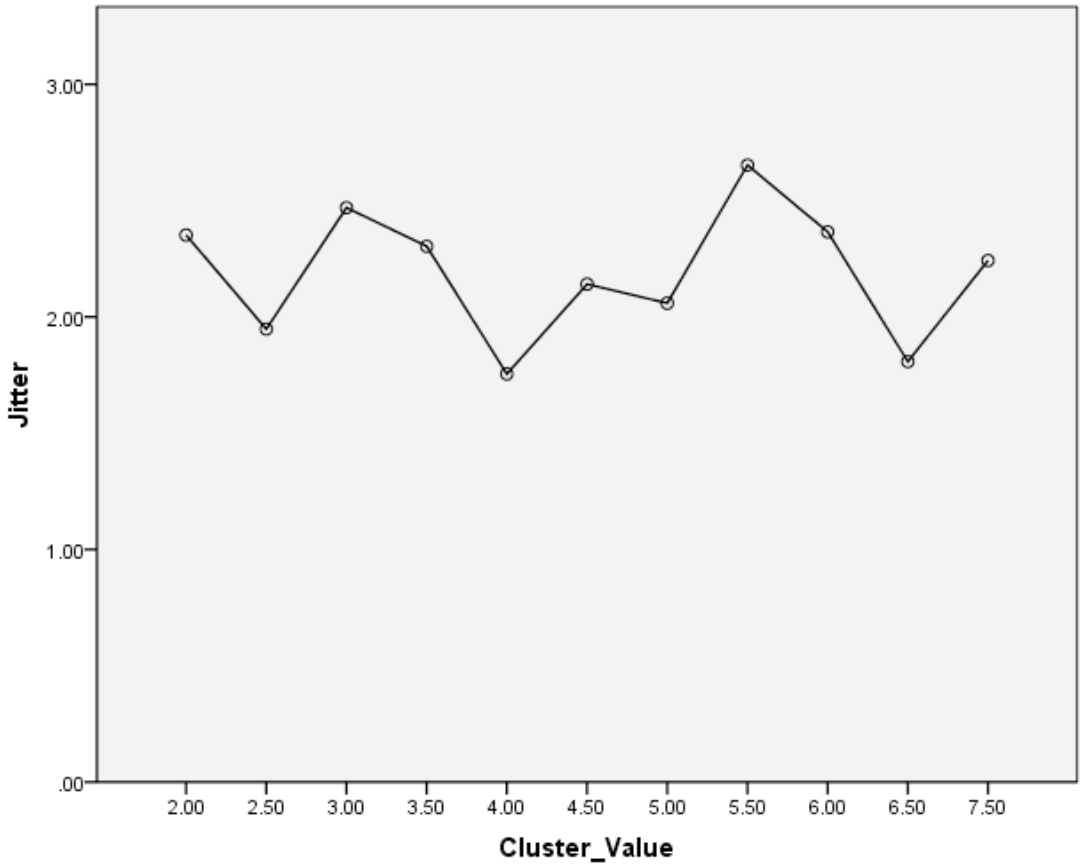
			Cluster_Value	Speech_Rate
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	.255
		Sig. (2-tailed)	.	.450
		N	11	11
	Speech_Rate	Correlation Coefficient	.255	1.000
		Sig. (2-tailed)	.450	.
		N	11	11



Jitter

Correlations

			Cluster_Value	Jitter
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	-.055
		Sig. (2-tailed)	.	.873
		N	11	11
	Jitter	Correlation Coefficient	-.055	1.000
		Sig. (2-tailed)	.873	.
		N	11	11



Shimmer

Correlations

			Cluster_Value	Shimmer
Spearman's rho	Cluster_Value	Correlation Coefficient	1.000	-.091
		Sig. (2-tailed)	.	.790
		N	11	11
	Shimmer	Correlation Coefficient	-.091	1.000
		Sig. (2-tailed)	.790	.
		N	11	11