A cross-cultural investigation of the vocal correlates of emotion

Alison A. Tickle

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

School of Education, Communication and Language Sciences
Newcastle University

March 2015

Declaration of originality

The material presented in this thesis is the original work of the candidate except as otherwise acknowledged. It has not been submitted previously in part or whole, for any award, at any university, at any other time.

Abstract

Universal and culture-specific properties of the vocal communication of human emotion are investigated in this balanced study focussing on encoding and decoding of Happy, Sad, Angry, Fearful and Calm by English and Japanese participants (eight female encoders for each culture, and eight female and eight male decoders for each culture). Previous methodologies and findings are compared. This investigation is novel in the design of symmetrical procedures to facilitate cross-cultural comparison of results of decoding tests and acoustic analysis; a simulation/self-induction method was used in which participants from both cultures produced, as far as possible, the same pseudo-utterances.

All emotions were distinguished beyond chance irrespective of culture, except for Japanese participants' decoding of English Fearful, which was decoded at a level borderline with chance. Angry and Sad were well-recognised, both in-group and cross-culturally and Happy was identified well in-group. Confusions between emotions tended to follow dimensional lines of arousal or valence. Acoustic analysis found significant distinctions between all emotions for each culture, except between the two low arousal emotions Sad and Calm.

Evidence of 'In-Group Advantage' was found for English decoding of Happy, Fearful and Calm and for Japanese decoding of Happy; there is support for previous evidence of East/West cultural differences in display rules. A novel concept is suggested for the finding that Japanese decoders identified Happy, Sad and Angry more reliably from English than from Japanese expressions. Whilst duration, fundamental frequency and intensity all contributed to distinctions between emotions for English, only measures of fundamental frequency were found to significantly distinguish emotions in Japanese. Acoustic cues tended to be less salient in Japanese than in English when compared to expected cues for high and low arousal emotions. In addition, new evidence was found of cross-cultural influence of vowel quality upon emotion recognition.

Early versions of a portion of this work were presented at conferences and published in proceedings:

Tickle, A., (1999). Cross-language vocalisation of emotion: methodological issues. In: J.J. Ohala, Y, Hasegawa, M. Ohala, D. Granville, & A.C. Bailey (Eds.) *Proceedings of 14th International Congress of Phonetic Sciences (ICPhS),* August 1-8, 1999, San Francisco, USA, 305-308.

Tickle, A. (2000). English and Japanese speakers' emotion vocalisation and recognition: A comparison highlighting vowel quality. In: R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.), *Proceedings of ISCA Workshop on Speech and Emotion.* September 5-7, 2000, Newcastle, Northern Ireland: International Speech Communication Association, 104-109.

To my parents, and to Nana

Acknowledgements

My thanks go to my supervisor, Gerry Docherty for all of his help, support and encouragement during the writing of this thesis.

I should also like to thank Barry Heselwood of Leeds University for his support and great sense of humour and am grateful to Leendert Plug, also of Leeds University.

This thesis would not have been possible without the kind voluntary participation of Japanese and English people who gave up their time to take part in the experiments conducted for this study. My gratitude also goes to all those at Newcastle University and INTO Newcastle University who have given me support in diverse ways over the years spent writing this thesis. Thanks to David Clark from Newcastle University Music Department for early conversations on the theme of this thesis. I am grateful to Simon Kometa for helpful advice on the statistical analysis in this thesis and to Doug Cudmore and Chris Letts for technical advice.

I should also like to thank previous colleagues at Barcelona University for their encouragement when I first had the idea for this thesis and to Albert for the conversations over trifásicas on the Rambla. Thanks to the musicians who have helped me escape when I needed to and to return with inspiration.

The reliable support of friends has been heart-warming. I am grateful to Lynne for being there over many years – and no, I doubt they never did go shopping! Thanks very much to Jadwiga Billewicz, Lynne Pickles, Jo Craggs, Joan Tickle and Dave Porthouse for their proof-reading assistance; it is very much appreciated. Any remaining errors are of course my own. There are bottles in the fridge and on the mantelpiece.

It seems a long time since the early days of discussions around the area of emotion in speech, in a sushi bar and various other venues in San Francisco. However, those conversations provided invaluable lasting encouragement for the writing of this thesis. My appreciation goes to Sylvie Mozziconacci, Marc Schröder, Ellen Douglas Cowie, Roddy Cowie, Veronique Auberge, Louis ten Bosch and Olivier Piot, and to all those working within the area of emotion in speech with whom I have had conversations, including those at the distant ISCA Conference in Newcastle, Northern Ireland, where many ideas on emotion in speech began to germinate.

Special thanks go to my parents, Joan and Ron Tickle for their unwavering love and support. I am grateful to all of my family for their stalwart support: to Annie Reynolds, who was and is always there; and in no particular order, to May and Roy for their considerable practical and moral support; to Audrey and Jack for helping me to keep my feet on the ground; to Vince and Sarah for their support and their haven of escape on Lake Geneva; to Joe, Hannah, Jack and Sam (in order of age!) for their sense of fun and for helping me to retain a sane perspective whilst writing this thesis; to Alan, Thomas, Holly and Oscar for the early morning walks which geared me up for writing and to all of my friends and family for their many kinds of support, all invaluable and much appreciated. Thanks in particular to Dave Porthouse for his love, help and support.

Whilst this is an objective study, as far as any human observer can make an objective study, I feel I have experienced all of the emotions investigated here during the course of completing this thesis and many others besides.

A conglomeration of inspiration has contributed to the writing of this thesis. Thanks to all those who were there.

As soon as human beings began to make systematic observations about one another's languages, they were probably impressed by the paradox that all languages are in some fundamental sense one and the same, and yet they are also strikingly different from one another.

> Ferguson, C. A. (1978). Historical background of universals research. In J.H. Greenberg, C.A. Ferguson, E.A. Moravcsik (Eds.) *Universals of human language, Vol. One: Method and theory*, 7–33. Stanford, CA: Stanford University Press. p.9.

Contents

Appendices

List of Tables

List of Figures

Introduction

> It is to be hoped that in addition to the recent trend to focus on naturally occurring expression in the field, there will also be more theoretically motivated and experimentally controlled studies—particularly across cultures—that will advance further our understanding of the psychobiological and sociocultural factors and mechanisms underlying the expression of emotion (Scherer, K.R., Clark-Polner, E. & Mortillaro, M., 2011, p.427).

1.1 Background and aims

The field of emotion is a complex area which has long inspired interest within the disciplines of philosophy, religion and the arts, including music. Since Darwin's seminal work *The Expression of Emotion in Man and Animals* (1872), emotion has in addition become an important area of research within many other areas including psychology, clinical psychology, psycho-physiology, evolutionary biology, ethology, sociology, ethnology, affective computing, child language acquisition, speech and language communication disorders and cross-cultural communication. Many theories of the nature and function of emotion have emerged, particularly over the past 80 years.

It was not until the second half of the twentieth century that there was much scientific interest in emotion since the area was regarded as somewhat vague and not seen to be measurable. However, there has recently been a considerable rise in interest in the study of emotion within neuroscience and psychophysiology, largely due to the development of instrumental techniques, such as fMRI scans, which can measure physiological as well as neurological states and responses of the human body including the brain. There has been considerable scientific research, particularly by Ekman and co-authors (see references below), into the facial expression of emotion and there has been a growing interest in the study of the vocal correlates of human emotion, the present study being one example.

Applications of such research are broad, including greater understanding of human communication and clinical applications. For example, within the area of human-machine interaction, the search for accurate machine-generated human emotion expression and recognition software has led to a considerable body of research, such as

the HAMLET project (Murray & Arnott 1993;1995; 2008) and the HUMAINE (Human Machine Network Interaction on Emotions) project which have brought together researchers from different disciplines. For an overview of the range of areas involved in the HUMAINE project see Petta et al. (2011). This project also discusses the ethical considerations of applications of findings from research into emotion (Cowie, 2011).

The nature/nurture debate remains one of the fundamental challenges facing science. The whole question of what emotions are and how they are expressed in various modalities, forms part of this question. Phonetics connects very closely with this as vocal expression is a key modality. One means of addressing the question of the extent to which emotional expression is psycho-physiologically or culturally determined is to conduct cross-cultural research. Whilst there is considerable research into possible cross-cultural and universal influences on the facial expression of emotion, there has been a lack of focus on the communication of emotion in speech until recently. Substantial mono-cultural research has been conducted in this area over the past two decades. However, at least partly due to methodological difficulties, few cross-cultural studies have been conducted so the potential contribution of the phonetics of emotion to the nature/nurture debate remains an area in need of considerable investigation.

Juslin and Laukka (2003) gave a comprehensive review of studies of emotion vocalisation in speech and in music. Of 104 speech studies, only 12 were cross-cultural in the sense that subjects from more than one culture either encoded or decoded vocal expressions of emotion. Of these, only 5 studies included any acoustic analysis of the data. More recent studies of cross-cultural vocal communication of emotion tend to include either encoders or decoders from a single culture and often no acoustic analysis of the data is conducted.

The key aim of this study is to contribute to evidence of possible pan-cultural and cultural influence on the vocal cues of emotion by highlighting examples of vocal expressions of emotion which are most reliably recognised both within the same culture and across unrelated cultural groups (English and Japanese) and by then analysing the vocal correlates of these expressions. This study therefore includes data encoded by both cultural groups and both in-group and cross-cultural decoding data. The most reliably decoded data was analysed acoustically.

The theoretical underpinnings of research in this area are intrinsically multidisciplinary. Scherer (2003), Spackman, Brown, and Otto (2009) and Ogarkova, Borgeaud and Scherer (2009) suggest that in order to advance research in this area, investigators need to be more aware of the discussion around the nature of emotions and the mechanism which underlies the vocal cues found. Scherer (2003) argues that "One of the major shortcomings of the encoding studies conducted to date has been the lack of theoretical grounding, most studies being motivated exclusively by the empirical detection of acoustic differences between emotions." (Scherer, 2003, p.234).

Chapter One therefore makes explicit the theoretical underpinnings of the debate surrounding the nature of emotion and its expression, highlighting the big issues around emotion and the major conundrums around the influence of nature and nurture as a necessary context for this thesis. Scherer et al. (2001) suggested that the investigation of the vocal expression of emotion can be particularly helpful in the search for universals in the expression of emotion, "…given its likely roots in nonhuman primate vocalisations and the extraordinary variability of language and communication systems that have evolved in different cultures" (p.90).

The term cross-cultural rather than cross-language is used in this thesis since it focusses on potential cross-cultural influences on the vocal expression of emotion. The languages spoken by the participants are used as a proxy for cultural diversity. Speakers of Japanese and English were chosen for comparison because they represent two very distinct cultures.

Humans have the potential to express emotion through verbal and non-verbal modalities. Non-verbal expression may include vocal, verbal, facial and/or gestural expression, and will tend to include a combination of these modalities in the expression of emotion. This thesis focusses on the investigation of cross-cultural similarities and differences in the vocal communication of emotion.

There is considerable evidence of universals in the facial expression of what are termed 'basic emotions' and the present study, like most previous cross-cultural research in this area, uses this evidence as a base to investigate the extent to which basic emotion cues are pan-cultural or culture-specific in the vocal mode of communication. The concept of basic emotions is also therefore explored in Chapter One.

Finding a rigorous, ethical and fruitful method of data collection remains a challenge particularly for cross-cultural research in this area. A simulation/self-induction method was constructed for the present study and the theoretical basis and rationale behind self-induction methods in general is discussed in Chapter One. The specific application of this method to the present study is explained in Chapter Five.

Most mono-cultural studies of the vocal expression of emotion have focussed on investigation of the role of one or more parameters related to fundamental frequency, intensity and duration in the expression of emotion, which are associated with the major prosodic parameters of pitch, loudness and timing. Following on from these studies, the few cross-cultural studies of the vocal expression of emotion which have included acoustic analysis of data from more than one culture have tended to include a very small number of parameters and to investigate the vocal communication of emotion in mainly Western cultures. The present study includes measurement of 11 acoustic parameters, which relate to fundamental frequency, intensity and duration, in the vocal expressions of emotion by speakers from one Western and one Eastern culture. This is an unusually large number of parameters to include in a cross-cultural study. This is not to say that other parameters such as voice quality and fundamental frequency contour are not relevant. However, as discussed further in Chapter Two, they are beyond the scope of this investigation.

Issues of validity are particularly relevant to such cross-cultural studies. For almost two decades, previous cross-cultural studies in this area have called for the development of a new 'balanced' methodology for encoding and decoding vocal cues of emotion in more than one culture, allowing more rigorous testing, comparing both In-Group and Cross-Cultural decoding of emotion expressed by each cultural group studied. One of the major aims of this study was to meet the challenge of developing a new balanced methodology which would test the validity of previous findings and could potentially highlight new evidence both in the decoding results and in the vocal cues recognised as expressing emotion in each culture. The present study aims not only to meet the challenge of constructing a new, balanced methodology, but also to create a 'symmetrical' methodology in which the testing method is as far as possible, the same for both cultures.

It is therefore vital to explore the methodological issues which this study attempts to address, with the goal of testing previous findings and possibly highlighting new

evidence emerging from this new balanced, symmetrical method of gathering cross-cultural data on the production (encoding) and recognition (decoding) of emotion in the voice. Given the importance of a new balanced, symmetrical method to other researchers in the field, the first section of Chapter Three is an issue-based discussion of the methodological issues and challenges highlighted in previous research. The second section of Chapter Three is a thematic explanation of the evidence and hints found in previous studies as to cross-cultural similarities and differences in the expression and recognition of emotion. These findings and hints led to the construction of the research questions, detailed in Chapter Four.

It is argued that this new, more rigorous methodology will potentially allow stronger confirmation of previous findings and may highlight evidence which has previously not emerged due to a lack of balance and symmetry in previous studies.

Since the vocal communication of emotion involves both a speaker and a listener, the inclusion of both in-group and cross-cultural encoding and decoding tests provides a strong evaluation of cross-cultural similarities and differences in the vocal cues they produce and recognise. There have been frequent calls for such 'balanced' studies (Galatà & Romito, 2010; Pell, Monetta, Paulmann & Kotz, 2009a; Pfitzinger, Amir, Mixdorff & Bösel, 2011; Scherer, Banse & Wallbott, 2001; Thompson & Balkwill, 2006). However, the even greater methodological difficulties implicit in balanced cross-cultural studies at least partly explains why there are so few balanced studies in this area. Pell et al. (2009a) comment:

> These methodological considerations likely explain why many researchers in the vocal literature have adopted ''unbalanced'' experimental designs which involve repetition of items produced by a single cultural group to a number of different listener groups… or presentation of vocal expressions produced by speakers of several different languages to listeners from a single language/culture… (p.109).

A key challenge, therefore, in this thesis was the construction of a rigorous, balanced methodology which addressed problems of translation and the trade-off between authenticity of the data and the need for consistency for comparative analysis across the two cultures. One of the novel aspects of this study is the symmetrical design of its encoding and decoding procedures, including the use of, as near as possible, the same pseudo-utterances for each culture. In addition, these were constructed to test for similarities and differences in the possible influence of vowel quality on the vocal

communication of emotion, which constituted a further aim of the present study. This study can be described as having symmetrical experimental procedures in that the same simulation/self-induction techniques are used to encode emotion in both cultures. It is one of the few cross-cultural studies in this area to include acoustic analysis.

The methodology of this study was designed with the goal of facilitating the further investigation of questions which have arisen in previous research regarding cross-cultural similarities and differences, 'In-Group Advantage' and the influence of vowel quality.

In their review of 97 studies of visual and prosodic emotion cues, Elfenbein and Ambady (2002) found that "…although emotions are recognised at above chance levels across cultures, there is also cross-cultural variation in recognition accuracy" (p.204). These authors used the term In-Group Advantage to describe this phenomenon, later formulating this as the Cultural Proximity Hypothesis (Elfenbein & Ambady, 2003), which states that the more related a decoder's language is to the encoder's language, the more accurately they decode the emotions encoded. According to this hypothesis, for example, cultures sharing cultural elements such as degree of individualism or collectivism or type of power structure would more accurately decode each other's expressions of emotion.

1.2 Structure

Following this Introduction, Chapters One, Two and Three provide the background of the present investigation, demonstrating the context and motivation for the study and reviewing previous research. Chapter One gives an overview of the theoretical, historical and multidisciplinary context of the thesis. Chapter Two describes the acoustic parameters measured in previous research in this area, including briefly highlighting issues surrounding the measurement of fundamental frequency, and discusses previous evidence found of vocal cues of emotion in both mono-cultural studies and in the few previous cross-cultural studies conducted in this area. Chapter Three explores the challenging and complex methodological issues involved in research, particularly cross-cultural research, into the vocal expression of emotion. Findings of previous studies are also presented here. Four research questions were formulated on the basis of findings and questions arising from previous research. All of

these questions relate to the aim of contributing to evidence of possible pan-cultural influence on the vocal cues of emotion by highlighting examples of vocal expressions of emotion which are most reliably recognised both within the same culture and across unrelated cultural groups (English and Japanese) and by then analysing the vocal correlates of these expressions. The first of these questions focusses on in-group decoding of emotion. The second question addresses cross-cultural decoding of emotion. Both these questions include consideration of whether some emotions are easier than others to decode and, if this is the case, whether the same emotions are more reliably decoded than others cross-culturally. Question three was constructed to test for In-Group Advantage and the fourth question tests for the influence of vowel quality on emotion decoding. These questions are presented in detail in Chapter Four.

A major aim of this thesis was therefore also to meet the challenge of constructing a new, balanced, symmetrical methodology for encoding and decoding experiments in cross-cultural research in this area with the purpose of addressing the research questions of this thesis, which arise from gaps and questions raised in previous research. Chapter Five explains this methodology and the rationale behind the encoding and decoding experiments. Materials used for these experiments are given in Appendices A, B, C and D.

Chapter Six presents and discusses the results of the decoding experiment. Chapter Seven explains the acoustic parameters measured and details the results of acoustic analysis, with some reference to the implications of evidence from acoustic analysis for decoding test results. Chapter Eight draws together the results of this study, exploring the implications of these results for the research questions of this investigation, and reflecting on how the findings of this research have informed the theoretical base of the present study. Limitations of the study are indicated and recommendations for future research are highlighted.

Chapter One

The nature of emotion - theoretical overview

> Everyone knows what an emotion is, until asked to give a definition (Fehr & Russell, 1984, p.464).

1.1 Introduction

The purpose of this chapter is to provide an overview of findings and theories relevant to the present study, focussing in particular on the debate surrounding possible universal and cultural influence upon the expression of emotion. The chapter aims to identify some of the relevant threads as applied to the expression of emotion in general and then to focus in on how that applies specifically to the vocal expression of emotion. It is not relevant to conduct a general exploration of multidisciplinary findings and theories of emotion here, but just to focus on those elements which are related to the vocal correlates of emotion. The present study aims to contribute to the debate surrounding universal and cross-cultural influence on emotional expression by testing for cross-cultural similarities and differences in the production and recognition of vocal cues of basic emotions by native speakers of unrelated languages. This chapter therefore briefly explores the parameters of this debate, including discussion of what is understood by the term 'basic emotion'.

It is argued that where phylogenetic continuity exists due to innate psycho-physiological response mechanisms, the characteristics of acoustic cues influenced by these response mechanisms is likely to be universal. Whilst there may be some emotions that are quasi-universal, there are also aspects of emotions which appear to be learnt and characteristic of particular cultures and even particular individuals and both of these ideally need to be factored into models of the expression of emotion.

In addition, other related elements which are specifically relevant to the encoding methodology constructed for this study are also introduced, including facial expression research, the 'facial feedback hypothesis', 'emotional contagion', affect bursts and the possible cultural bias in emotion words used in empirical investigation of the expression and perception of emotion (Wierzbicka, 1992).

1.2 Universal versus social constructivist theories of emotion

The extent to which human emotion is universally or culturally determined has been a subject for debate since Darwin's work, *The Expression of Emotion in Man and Animals* (1998). This possibly sounds like an unusual date for this reference since the original work was published in 1872. However, this edition includes useful comment by Ekman, who found strong empirical evidence for the existence of quasi-universal facial expressions of certain emotions as discussed below. The present study contributes to this debate by its focus on cultural and cross-cultural influence on the vocal expression of emotion. Darwin (1998) found evidence of universals and phylogenetic continuity in the expression of emotion in humans. He suggested that anger was signalled visually by dilated nostrils, compressed mouth, furrowed brow, wide open eyes, erect head, expanded chest, arms held rigid to the sides of the body, stamping the ground, the body swaying backwards and forwards and trembling and was signalled vocally by a loud voice and high pitch. A linguistic signal can be verbal or non-verbal and is anything which transmits meaning. He suggested that these outward expressions arose from internal physiological changes such as increased heart rate, faster respiration and muscle tension. The psycho-physiological influence upon the expression of emotion is discussed further below. However, Darwin also commented on the importance of the expression of emotion for social communication. Whilst he focussed upon innate, universal aspects, he does not suggest that cultural influence upon emotion does not exist.

The idea that emotions are innate was challenged in the twentieth century. From 1928 (Mead, 1928) onwards, anthropologists presented evidence which they claimed refuted Darwin's conclusions on the innateness of emotion. Behavioural psychologists since Skinner (1935) have focussed on investigation of the role of environment and response in human emotion, claiming that it is learning which influences what we are as humans and how we behave. These studies and work by Tomkins (1962; 1963) on basic human emotions made a considerable contribution to fuelling the nature versus nurture debate which continues today. Empirical studies of the production and recognition of emotion provide some of the strongest evidence on both sides.

Early cross-cultural studies of facial expression (Ekman, 1972; Ekman et al., 1969; Ekman & Friesen, 1971; Ekman et al., 1972; Izard, 1971; Matsumoto, 1996) demonstrated strong evidence for the existence of universal basic emotions including

happiness, sadness, anger, fear, surprise and disgust, thus appearing to confirm Darwin's observations. See Section 1.3 for further explanation of the theory of basic emotions. Despite this evidence, in 1980, Averill presented a constructivist view of emotion, arguing that human emotions are cultural creations, based on nurture rather than nature.

Ekman and Friesen (1971) had argued that just because strong evidence for innate influence upon the expression of emotion had been found, this did not mean that there was no cultural influence. They suggested that there is cultural influence upon the circumstances which may elicit a particular emotion and the resulting action performed by an individual. They also indicated that display rules of different cultures will influence the facial expression of emotion. Display rules may be defined as culturally prescribed norms about which emotions are appropriate to express when and where and how these emotions should be expressed. Matsumoto (2006) reported work by Friesen (1972) and Ekman (1972) which induced basic emotions in Japanese and North American participants and found that whilst the North American group tended to produce facial expressions of basic emotions whether alone or not, Japanese participants produced similar facial expressions to those of the North American group when alone, however, they tended to smile more when not alone. They conjectured that Japan has a collectivist culture which focusses on not standing out from the crowd whilst the American culture is individualist, individual autonomy being seen as particularly important. The conclusion was drawn that evidence was found for universals in facial expressions of emotion when alone, which were affected by cultural display rules when not alone. For a recent summary of critiques of this view, see Leys (2010).

The current online Oxford English Dictionary (Oxford Dictionaries, 2015) defines an emotion as 'A strong feeling deriving from one's circumstances, mood, or relationships with others' and the mass noun as 'Instinctive or intuitive feeling as distinguished from reasoning or knowledge'. However, in cross-cultural research it is important to bear in mind that the conception of what an emotion is may vary from culture to culture (Wierzbicka, 1992; 2009). In Tibetan, for example, there is no general word for "emotion". Ekman, Davidson, Ricard and Wallace (2005) argue that this is consistent with the current scientific view of the anatomy of the brain, that cognition and affect are processes within the neurological circuitry which are vitally intertwined with each other and cannot be separated into distinct processes. Within the psychology literature, emotion tends to be conceptualised as a complex state of feeling that results in and/or

from physical (including neurological) and psychological changes which influence thought and behaviour.

Whilst western cultures tend to place considerable focus on the sense of externalising or expressing something, there may be less of a focus on this aspect in many Eastern cultures where a sense of the need to control the internal process of emotion may well cause speakers to inhibit acoustic cues more. A fundamental concept in Hindu tradition is that of 'Chitta vritti', which roughly translates as fluctuations of consciousness (Woods, 2003). It is argued that these fluctuations, due to emotions and thoughts, need to be controlled in order to achieve stillness and clarity of mind. The focus is therefore upon the internal process of emotion caused by perception of something regarded as external, in contrast to the western interpretation of externalising or expressing something.

In terms of specific emotions, even where there appears to be a direct translation of the word used for a particular emotion in different cultures, this does not mean that the experience of the emotion is necessarily the same and this may impact upon its expression. For example, Mesquita and Karasawa (2002) argued that whilst the experience of 'Happy' by American English speakers is focussed upon self-esteem and self-control, the sense of connection with others is also particularly relevant for Japanese speakers experiencing 'Happy'.

The debate around universal and cultural influence upon human emotion continues between evolutionary psychologists and social constructivists. Research in this area often focusses on basic emotions, one characteristic of which is argued to be their universality. The present study investigates the vocal correlates of four basic emotions. An exploration of what is understood by this term is given below.

1.3 Basic emotions

Emotion has been conceptualised in various ways. However, it is often argued (Darwin, 1998; Ekman, 1972; 1992; Ekman, Sorensen & Friesen, 1969; Izard, 1971) that there exist universal 'basic emotions'. Basic emotion is a notion which occurs across a number of domains in which emotion is investigated, in fields as diverse as psychology (including clinical psychology), neurophysiology, psychophysiology, evolutionary

biology, zoology, human-machine interaction, bilingualism, cross-cultural communication, child language acquisition, facial expression, gesture and vocal expression. The notion of discrete forms of expression associated with distinct basic emotions is also highly evident in previous investigations of human expression of emotion (see Cowie & Cornelius, 2003), particularly cross-cultural studies of the expression and recognition of emotion, where evidence of universal influence upon the expression of emotion is most likely to be found. The strongest evidence for this universality has come from cross-cultural research into the facial expression of emotion (Ekman & Friesen 1971; Izard, 1971; Matsumoto 1996), as discussed above. Ekman (1972) found evidence supporting basic emotion theory in his research into universals in the facial expression of six specific basic emotions: happiness, sadness, anger, fear, surprise and disgust, the strongest evidence being found for the first four of these, which are the four emotions investigated in the present study. Surprise and disgust were often confused (Ekman 1972). Contempt was added to this list of basic emotions in a study by Ekman and Friesen (1976).

However, Ekman (1999) refers to universal "emotion families" rather than specific universal emotions, clarifying the idea that what may be seen as basic, innate emotions may be adapted to different cultures in different ways. Emotions such as contentment and melancholy, for example, have sometimes been referred to as 'social' or 'secondary' emotions (Murray & Arnott, 1993). According to research in neuropsychology (Zillmer, Spiers & Culbertson, 2008), whilst basic emotions are viewed as innate, sensory experiences, 'social' emotions, such as pride, shame and embarrassment, are acquired through learning and socialising. This appears to align with a constructionist view that there are many different kinds of happiness or anger or sadness and so on. Shaver, Schwartz, Kirson, & O'Connor (1987) found that emotion terms tend to be grouped together in clusters which appear to correspond to basic emotions. Prototype systems have been proposed in which each basic emotion prototype includes other emotions. For example, contentment and pride may be included within the joy prototype.

Whilst there continues to be controversy as to the nature and function of emotion, studies such as Shaver et al. (1987) have shown that there is fairly general consensus on which of a long list of psychological states, are good examples of emotions and which are not. Cowie and Cornelius (2003) discuss suggestions made for such lists and propose a list of emotion terms relevant to emotions in everyday life. 16 emotion terms emerged from research by Cowie et al. (1999), which include Angry, Afraid, Sad,

Happy, Worried, Amused, Pleased, Content, Interested, Excited, Bored, Relaxed, Disappointed, Confident, Loving and Affectionate. These authors also give examples of previous lists (Cowie et al., 1999, Table 2).

Prinz (2004) argued that "...every emotion that we have a word for bears the mark of both nature and nurture. Each is built up from a biologically basic emotion, but its conditions of elicitation, and hence its content, is influenced by learning" (2004, p.16). Wierzbicka (1992; 2009) criticised research claiming to find evidence of universal expression of basic emotions on the grounds of cultural bias in the terms used to represent these emotions. In defence of his findings, Ekman (Darwin, 1998) pointed out that studies (Izard, 1971; Ekman & Friesen, 1971; Ekman, 1972) had also been conducted without the use of emotion words and these also demonstrated the existence of universals.

Ekman (1992) argued that basic emotions have nine characteristics in common, all of which have a 'biological contribution'. The first characteristic, that a basic emotion has **distinctive universal signals**, is the focus of the present investigation. To be defined as a basic emotion, Ekman (1992) suggests that the emotion should also be **present in other primates**; this relates to the discussion below of phylogenetic continuity. He argued that basic emotions have a **distinctive physiology**, based on the activity of the autonomic nervous system (ANS), also including activity in the central nervous system and that that basic emotions demonstrate **coherence between this physiological response and display systems**, including facial, vocal and gestural expression. This is discussed further below. He suggests that basic emotions are **triggered by distinctive events which are universal**. For example, an event perceived as threatening will lead to a feeling of fear. Appraisals of events which lead to the experience of basic emotions are, he argues, **quick and automatic** and basic emotions have a **quick onset** and a **brief duration**. Finally, Ekman (1992) argues that basic emotions have an **unbidden occurrence**; in other words "One can not simply elect when to have which emotion" (p.189).

Many previous studies now provide evidence of the phylogenetic continuity of basic emotions. The first empirical evidence presented evidence to suggest that universal, phylogenetically continuous, psycho-physiological mechanisms were used in the facial and vocal expression of basic emotions (Darwin, 1998). Relatively recent evidence (Papousek et al., 1992) suggested that of all nonverbal channels, vocal expression may

be the most phylogenetically continuous and is therefore a universal device for the communication of emotion. Panksepp (1998) also presented more evidence of phylogenetic continuity of basic emotions. A study by Jurgens and Hammerschmidt (2006) compared acoustic cues in the vocal expression of emotion by monkeys and humans and concluded that

> …monkey and man differentiate aversive from non-aversive emotional states vocally in a very similar way. This suggests that the vocal expression of emotional states in humans and monkeys is homologous in the strict sense. As that branch of the phylogenetic tree leading to the squirrel monkey already separated from that branch leading to modern humans about 45 million years ago, we may conclude that the phylogenetic roots of human emotional vocal behaviour reach back at least 45 million years. (Jurgens & Hammerschmidt, 2006. p.6)

It has been claimed (Krumhuber & Scherer, 2011; Scherer, 1994) that the expression of emotion, particularly basic emotions, in both speech and music may have developed from primitive interjections or affect bursts. Scherer (1994) defined affect bursts as "very brief, discrete, nonverbal expressions of affect in both face and voice..." (p.170). A "gestural component" was added to the definition in Krumhuber & Scherer (2011). Heylen et al. (2011) comment that affect bursts can have various vocal forms:

> ...from non-phonemic vocalisations such as laughter or a rapid intake of breath, via phonemic vocalisations such as [a] or [m], where prosody and voice quality are crucial in conveying an emotion, to quasi-verbal interjections such as English 'yuck' or 'yippee' for which the se9gmental form transports the emotional meaning independently of the prosody. (p.334)

Schröder et al. (2006) tested cross-cultural recognition of affect bursts by Dutch and German listeners of emotion expressed in affect bursts produced by native speakers of German. Whilst consideration needs to be made for the fact that German and Dutch are related languages and that the cultures are similar and geographically close, this study found evidence of considerable cross-cultural recognition. Where Dutch listener recognition was less accurate the authors concluded that the German listeners recognised the affect from the language-specific segmental form.

In 1872, Darwin suggested that the use of certain sounds by animals could signal specific basic emotions and that there may be phylogenetic continuity in human use of certain vowel sounds. For example, he argued that that the technique of vocal tract lengthening was used by certain animals to signal a larger sound source when expressing anger or aggression and that [u], with its lengthening of the vocal tract, may be more

likely to suggest aggression or anger in human vocal expression of emotion. The end of the twentieth century saw renewed interest in the acoustic features of emotional signals in mammals. Research by Ohala (1984) also suggests that animals use a 'frequency code' to signal, for example, larger size, strength and aggression. There is little recent research on this; however, the study by Xu and Chuenwattanapranithi (2007) supports this idea. This aspect is relevant to the investigation of the influence of vowel quality on recognition of emotion in the present study and is discussed further in 3.9.6.

Physiological signs of emotion which Darwin (1998) reported observing as common across species, included, for example, effects upon heart rate, respiration, perspiration, salivation, pupil and nostril dilation, facial expression and the muscular system in general which also led to effects upon the voice. (Darwin, 1998) Ekman et al. (1983) suggested that there are distinct patterns of ANS activity for different emotions and Ekman (1992) argued that there is cross-cultural consistency in ANS patterns for specific emotions.

There is also neurological evidence which supports the concept of basic emotions. There is evidence of a differential neurological response to different basic emotional states. For example, Le Doux (1996) found that negative emotions, and particularly fear, activate the amygdala bilaterally. LeDoux *(*1996) suggested that each emotional function may be underpinned by a different neural system, each emotion having evolved for a different purpose, such as providing food, caring for offspring, finding mates and bonding with mates.

Posamentier and Abdi (2003), focussing on non-verbal visual signals and Wildgruber et al. (2006) focussing on auditory signals, review literature which identifies neural networks which are activated when perceiving non-verbal visual and auditory signals of emotion, particularly specific basic emotions. There are fewer studies on neural networks activated in multi-modal audiovisual non-verbal emotion communication (e.g. Ethofer et al., 2006; Kreifelts et al., 2007). There is also recent support in the literature for the existence of distinct patterns of activity within the auditory cortex, for the basic emotions happiness, sadness, anger, fear and disgust (Ethofer et al., 2009). However, there is often an implicit assumption in neurological research that where patterns of neural activity for basic emotions are found, these will be found cross-culturally; the author is unaware of any neurological studies which compare activity in the auditory cortex for different emotions across different cultures.

From the end of the twentieth century, there has been increased interest in the search for evidence of cross-cultural similarities in the vocal expression of basic emotions (Scherer, Banse, & Wallbott, 2001). The reasoning is that there are physiological responses to different emotional states which may have implications for the experience and expression of emotion. For example, Scherer (1986) argued that physiological signs such as a faster heartbeat and faster breathing occur with increased arousal, as in fear, anger and joy and that this would be expected to lead to faster speech. Increased muscle tension would tend to lead to higher pitch which has been found for anger, fear and joy and relaxed muscle would tend to lead to a lower pitch as found in previous studies for sadness and which we would expect to find for calmness. Reduced salivation would also be expected to influence the voice.

Banse and Scherer (1996) found monocultural evidence of emotion-specific vocal profiles. Scherer (1986) claimed that the acoustic parameters of these profiles are based on physiological changes which are emotion-specific. Both the ANS and somatic nervous system (SNS) can affect the speech production process. The activity of these two systems serves to maintain the body in a state of homeostasis (an optimal metabolic state). Reactions to emotion arousal constitute the body's attempts to return itself to this balanced state. The respiratory system, muscle tone and the production of saliva and mucous can all be affected by emotion arousal and can influence the voice. In terms of a multimodal response, Scherer & Ellgring (2007) predicted that "...an anger-producing event will generally produce a multimodal configuration consisting of knitted brows, clenched teeth, raised fist, body leaning forward, and a loud and strident voice" (p.159).

This section has discussed the concept of basic emotions, including potentially universal involuntary phylogenetically continuous physiological effects on the expression of emotion. The following section compares these effects to those of cultural 'display rules'.

1.4 Innate and cultural effects upon the psycho-physiological processes influencing vocal correlates of emotion

Scherer, Ladd and Silverman (1984) focussed upon how this conceptual space around emotion, discussed above, applies to the vocal correlates of emotion. They developed

the covariational / configurational model to distinguish involuntary physiological effects from the effects of cultural 'display rules' upon the vocal correlates of emotion. According to this model, involuntary physiological effects are known as covariational or 'push' effects, and include effects upon the vocal mechanism, such as influence upon prosody. For example, Scherer, Ladd and Silverman (1984, Vol.3) suggest a "direct covariation between the amount of muscle tension increase measured by electromyography and the increase in fundamental frequency" (p.91). Scherer (1986) gives the example of evaluation of a stimulus as dangerous and necessitating action. He argues that in this situation, f0 would increase due to increased muscle tension and salivation would also be reduced. This would result in a higher pitched voice.

There has long been interest in understanding involuntary physiological effects such as muscle contraction, faster or slower heartbeat, respiratory effects, raised or lowered temperature, sweat production and increased or reduced salivary secretions which are now viewed as effects within the ANS. These ANS effects are common to humans whatever their language or culture. The covariation model presented by Scherer (1986) incorporated these physiological influences in a theoretical model which explained how vocal expressions of emotion could be pan-cultural. He suggested that there are involuntary emotion-specific physiological changes which are linked to changes in acoustic patterns, which are in turn recognisable as expressing particular emotions. Scherer (1989) called for further research into the nature of these processes.

Neuroscientific research (e.g. Le Doux 1996) has argued that certain contexts are likely to produce similar physiological ANS effects on humans generally. Le Doux (1996) gives examples of emotions which may have evolved for different purposes including defending against danger, procuring food, finding or bonding with mates and suggests that these may be underpinned by specific neural systems.

These 'push' effects are distinguished from configurational or 'pull' effects which are caused by the influence of cultural 'display rules'. Different cultures may be more or less likely to attempt to inhibit or maybe even exaggerate psycho-physiological effects Johnstone, van Reekum and Scherer (2001). We would therefore expect that 'push' effects will tend to be similar cross-culturally whilst 'pull' effects will tend to differ. In addition, Scherer (1986) commented that "The configuration model seems to adequately describe the communication of more "cognitive" affect states such as "doubt", whereas the covariation model seems more adequate for basic emotions such as anger (1986

p.145). It is possible that cultural display rules may inhibit the expression of emotion or may demand an alternative form of expression from that governed by 'push' effects. Speakers may attempt to voluntarily influence their vocal mechanism in order to adjust their self-presentation. Differential cultural display rules may result in different voluntary effects cross-culturally. It is also possible that the effect on the vocal mechanism of attempts to reduce the salience of involuntary effects may cause further effects on vocal settings. If we extend the previous example of response to perception of a dangerous situation, whilst covariational effects would tend to lead to a rise in vocal pitch, the subject may attempt to control this, perhaps because a high pitched voice is seen as socially unacceptable in the subject's culture. A further effect upon the voice, suggested by Scherer (1986), is that that the voluntary force applied to the vocal mechanism in attempts to reduce f0 may cause harsh voice.

Cross-cultural investigation of emotion-specific vocal cues will help to indicate possible cross-cultural similarities and differences. Where similar vocal cues are used, this would imply similar psycho-physiological effects and pan-cultural tendencies.

The collection of psycho-physiological evidence of emotion-specific physiological responses on the vocal mechanism is not straightforward. Johnstone et al. (2001) suggest that "…specific combinations of appraisal responses might trigger coordinated physiological reactions that, in turn, lead to corresponding patterns, or profiles, of vocal parameters." (Johnstone et al., 2001, p.274). The Appraisal Model of Emotion (Arnold, 1960) focusses on the role of individual subjective evaluation or appraisal of events as a cause of individual emotional responses. Appraisal or evaluation by automatic cognitive processes is now considered a central concept in cognitive theories of emotion. Appraisal models of emotion focus on the importance of individual appraisal or evaluation in causing someone to experience emotion and appraisal may vary from culture to culture and from individual to individual. Scherer (1986) applied the appraisal model to research into human vocalisation of emotion in the form of the component process model. This model is discussed further in 3.6. Scherer (1986) defines emotion as "a series of interrelated adaptive changes in several organismic subsystems following antecedent events evaluated to be of major relevance to an organism's goals" (Scherer, 1986, p.146). Gerrig & Zimbardo's (2002) model of emotion includes these same components, defining emotion as "A complex pattern of changes, including physiological arousal, feelings, cognitive processes, and behavioural reactions, made in response to a situation perceived to be personally significant".

The potential role of individual personality in the vocal expression of emotion often goes unacknowledged. As found in recent studies in behavioural genetics (see e.g. Bouchard, 2004; Radulovic & Stankovic, 2007), individual personality is influenced by a complex interplay between genetics and environment. In previous literature investigating cross-cultural influence upon the vocal correlates of emotion, there is a general tendency to attribute innate covariational aspects to universal influence and configurational aspects to cultural influence. However, it is possible that different, genetically derived aspects of individuals differentially influence covariational psycho-physiological processes and that configurational psycho-physiological processes also demonstrate individual variation; both the conscious and unconscious mind may influence this variation and an individual's disposition to become conscious of, and to understand their own and others' emotions could also play a role. Individual variation is new territory worthy of future research but is beyond the scope of the present study.

In addition to appraisal of external events triggering psycho-physiological response mechanisms leading to the outward expression of emotion, it has been argued that expressions of emotion can themselves result in psycho-physiological response mechanisms, leading to initiation of or adjustment in the emotion experienced. This effect may occur due to adjustments in one's own facial, bodily or vocal expression, which relates to the James-Lange Theory (Lange, 1885) or due to observation of expressions of emotion in others, as in the Theory of Emotional Contagion (Hatfield et al., 1994). Both of these theories, relating to the induction of emotion, are discussed in more detail below, as they are relevant to the method used to encode emotion in the production experiment of the present study.

1.5 Induction of emotion

1.5.1 *The facial feedback hypothesis*

Many studies have included reference to the James-Lange theory (Lange, 1885), whether they support or challenge it, debate centring around which comes first, the physiological experience or the psychological experience of emotion. According to the James-Lange hypothesis, when a stimulus is received, such as a poignant memory, for example, the physiological reactions to the stimulus, such as muscle tension, raised heart rate, perspiration and dry mouth and the related vocal, facial and or bodily expressions

actually cause the feeling of the emotion. In other words, people feel happy because they smile and sad because they cry, whilst they feel angry because they frown and fearful because they flee from danger. Whilst Lange (1885) focussed on the visceral, autonomic domain, in 1890, James encompassed all bodily changes - visceral, muscular and cutaneous effects - including the face (James, 1983). This theory is still relevant today, relating to the facial feedback hypothesis discussed below.

The Cannon-Bard theory of emotion (Cannon, 1927) challenged the idea that emotional states follow visceral activity, using as evidence research which showed that even when the viscera were isolated from the central nervous system, emotional behaviour was still present claiming that the experience of emotion stimulates physiological responses rather than vice versa. For example, it was suggested that we can feel happy whether or not we smile.

There continues to be debate as to whether we feel happy because we smile and sad because we cry or do we smile because we are happy and cry because we are sad. It could be argued that the process may be continuous through feedback between the brain and physiological processes and expression and that perhaps the cycle of interaction between involuntary and cognitive processes in the brain and nervous system and vocal, facial and bodily expression can be triggered, or indeed consciously adjusted, at any point in the cycle.

Whilst Darwin (1998) found evidence of universals in the expression of emotion, he also referred to the human ability to control emotion:

> The free expression by outward signs of an emotion intensifies it. On the other hand, the repression, as far as this is possible, of all outward signs softens our emotions. He who gives way to violent gestures will increase his rage; he who does not control the signs of fear will experience fear in a greater degree; and he who remains passive when overwhelmed with grief loses his best chance of recovering elasticity of mind… Even the simulation of an emotion tends to arouse it in our minds (1998, pp.359-360).

This suggested that humans can intensify, reduce or self-induce emotion through specific physiological changes. The process is more complex when we consider that cultural display rules may also influence facial expressions and whether they are posed, intensified or inhibited (Matsumoto, 2006). The facial feedback hypothesis, which also directly relates to the James-Lange theory (Lange, 1885), suggests that facial expression can cause emotion experience. Distinctions have also been found between different

kinds of smile. 'Duchenne' smiles, which are genuine expressions of happiness, involve both the zygomatic major muscle, which causes the corners of the mouth to rise, and the orbicularis oculi muscle, which is used to raise the cheeks and cause wrinkling around the eyes. 'Non-Duchenne' or 'false' smiles are viewed more as social smiles and involve only the zygomatic muscles. Duchenne smiles may help to self-induce happiness, whilst Non-Duchenne appear not to have this effect (Messinger et al., 2001).

Hennenlotter et al. (2008) showed that imitating facial expressions of emotion could also trigger emotion-related neural activity: "…peripheral feedback from face muscles and skin during imitation modulates neural activity within central circuitries that are known to be involved in the representation of emotional states" (2008, p.5). This relates to the concept of emotional contagion discussed below. No conclusions are drawn regarding any effect of "...reduced facial feedback onto 'felt' emotions." (Hennenlotter et al., 2008, p.5). It is possible that a feedback effect may also be possible during the process of expression of emotion in the voice, although there is a lack of research in this area, just as there is relatively less research on emotion in the voice compared to the face in general.

1.5.2 *Theory of emotional contagion*

Apart from the role of an individual's expression in intensifying, reducing or self-inducing emotion in the individual, there has also been some interest in the effect of an individual's expression of emotion upon that of the observer. Aristotle (1991) made the suggestion that human vocal expression of emotion could affect the emotions of the listener.

The Singer–Schachter theory (Schachter & Singer, 1962, in Reisenzein, 1983), also known as the Two-factor theory, found evidence that subjects in the same physiological state (due to an injection of adrenaline) have different emotional reactions, expressing either anger or amusement depending on whether another person in the situation displays that emotion. This is related to the concept of 'Emotional Contagion'. Hatfield et al. (1994) claimed that emotion may be transferred from one person to another during communication due to the tendency to mimic facial expressions. They define primitive *emotional contagion* as "The tendency to automatically mimic and synchronize facial expressions, vocalisations, postures, and

movements with those of another person's and, consequently, to converge emotionally" (1994, p.5). Recent research in neuroscience also lends support to the existence of facial feedback and emotional contagion. Carr et al. (2003) and Lee et al. (2006) have demonstrated that the imitation of facial expressions can cause neural activation within the amygdala and other limbic regions. Wicker et al. (2003) found fmri evidence of the same area of the brain being activated when one is feeling disgust as when one observes another person feeling disgust.

There is also evidence that music can induce emotion in the listener (e.g. Webster & Weir, 2005; van der Zwaag et al., 2009; Pinchot-Kastner & Crowder, 1990). See 3.7.2.

1.6 The Brunswikian Lens Model applied to vocal correlates of emotion

Scherer (1986; 2003; 2013a) and Scherer et al. (2011) suggest that research in the area of the vocal communication of emotion should employ his adaptation of The Brunswikian Lens Model (Brunswik, 1956; Scherer, 1986; Scherer, 2003), arguing that this captures the whole process of the vocal communication of emotion. This model was originally used by Brunswik (1956) to describe visual perception and involves modeling the process through which a visual stimulus is created, transmitted and received. Scherer (1986; 2003) has taken the principles of the Brunswikian Lens Model and applied it to vocal communication.

Verbal, facial and gestural cues may also be present and influence how the decoder decodes the emotion of the encoder. However, Scherer's version of the Brunswikian Lens Model focusses specifically on vocal cues. As in other models of speech communication, like the speech chain model suggested by Denes & Pinson (1993), a connection is drawn between the encoder and decoder along a temporal continuum, involving the vocal apparatus of the encoder, the transmission of an acoustic signal and the decoder's hearing apparatus. However, in the Brunswikian Lens Model applied to the vocal correlates of emotion, the speaker's vocal apparatus is affected at a 'phenomenal level' by their reactions to an event. The physiological effects of these upon the vocal apparatus, whether covariational or configurational, result in acoustic cues ('distal cues'), such as characteristics of f0, which are transmitted to the listener (transmission phase, during which interference may occur). The decoder then perceives

the auditory cues ('proximal cues'), such as pitch characteristics, and makes a judgement on the emotional state of the speaker based on these cues. Within this framework, communication is facilitated by the potential for redundancy in the set of vocal cues which combine to communicate the emotional state of the speaker to the decoder. Where not all cues are available, the potential for multiple and overlapping cues allows the decoder to focus on remaining available vocal cues when making a judgement on the affective state of the speaker. Each of the stages (emotional state of the encoder, distal cues, transmission signal, proximal cues and attribution by the decoder) and the fit between them can be measured at an 'operational level'.

This model emphasises the importance of consideration of decoder identification of emotion targets as well as analysis of the vocal signal produced by the encoder. Whilst Scherer, has been calling for the use of his adapted version of The Brunswikian Lens Model in research in this area for the past 19 years (Scherer, 1986; Scherer 2003; Scherer et al. 2011; Scherer, 2013a), very few studies have analysed their decoding results as well as conducting acoustic analysis on the reliably decoded data.

Where cross-cultural studies are concerned, this approach is even more complex since both in-group and cross-cultural experiments and analysis need to be performed. Most previous cross-cultural studies are unbalanced studies, in that they do not include both in-group and cross-cultural decoding experiments, and most are recognition studies in that no acoustic analysis of the data is performed. Where acoustic analysis is performed, this has tended to include acoustic analysis of the data from a single culture. Scherer has often highlighted this as a problem since the full picture is not explored. See, for example, Scherer et al. (2011).

In line with Scherer's adaptation of The Brunswikian Lens Model (Scherer, 2003), a fuller understanding of the similarities and differences in the vocal correlates, or 'distal cues', of basic emotions across different cultures may therefore be gained by analysis of the results of recognition tests performed by subjects from each culture, including both in-group decoding (encoders and decoders from the same culture) and cross-cultural decoding (encoders and decoders from different cultures).

Analysis of the vocal correlates (distal cues) of reliably encoded (production) material for each culture is also required if a more complete picture is to be observed. Previous

studies have not tended to include all of these elements, as will be discussed further in Chapters Two and Three.

1.7 Summary

It has been argued by "psychobiologically minded researchers" (Scherer et al., 2001, p.89) that universals may exist in the vocal encoding of human emotion and that these may be "based on emotion-specific physiological patterning affecting voice production..." (Scherer et al., 2001, p.89). Those with social constructivist sympathies would emphasise the role of cultural dependency in the vocal expression of emotion. Zinken, Knoll and Panksepp (2008) comment: "It seems that the research traditions of being either interested primarily in universals or in diversity are so strong that they still frame the current debate." (p.7). However, combined investigation of both these influences will perhaps help to further knowledge in the area of human emotion, including understanding the characteristics of human vocal expression of emotion.

As discussed, there is some evidence of a possibly innate biology of emotion, which may be underpinned by neural circuits and interaction within the autonomic nervous system, leading to physiological effects which can affect the vocal mechanism. There is also evidence of distinct patterns of neurological activity for basic emotions, including within the auditory cortex (Ethofer et al., 2009). Evidence of cross-cultural recognition and similarities in acoustic correlates of basic emotions would provide evidence of acoustic correlates which may link to this neurological activity.

The theoretical base of the study focusses on the relative contribution of factors which may be quasi–universal versus culturally specific influence on the vocal expression of emotion. This study takes a step towards this goal by testing for recognition agreement and common acoustic characteristics in the vocal expression of emotion in two cultures whose difference is encapsulated in their unrelated languages, thus aiming to contribute to evidence of possibly universal characteristics and cultural differences in the vocal expression of emotion.

This chapter has then provided a very selective review of the literature on emotion. What this has shown is that all research and models tend to point to universal, physiological type characteristics which are innate and cross-species but that there are

also specific learnt or cultural aspects. This is around the interaction between person and context, between encoder and decoder and around the different ways in which emotions arise. The following chapter focusses more on the characteristics of the vocal signal produced for different emotions found in previous investigations.

Chapter Two.

The vocal correlates of emotion – previous production studies

2.1 Introduction

The previous chapter discussed what is understood by the term emotion, focussing in particular on basic emotions. A brief description was given of relevant research and models of emotion as background to the debate on possible innate and cultural influence upon the vocal expression of emotion. We also need to consider how we know which emotion is being measured. Of the few cross-cultural studies which have conducted acoustic analysis, most have included a decoding test to verify the reliability of the data. Where acoustic analysis has been included, this has tended to be conducted on data encoded by a single culture.

The range of acoustic parameters also needs consideration before discussing which cues have been found in previous investigations of the emotions studied here and whether these are potentially universal cues, possibly as a result of covariant psycho-physiological processes, as discussed in Chapter One, or due to the influence of cultural display rules.

According to Scherer (2003), the results of previous monocultural and cross-cultural studies which have looked at decoding emotion in the voice (Chapter Three) provide evidence that basic emotions can be decoded from vocal cues at a level beyond that predicted by chance, at around 55-65%. This suggests that acoustic features are present which allow decoders to distinguish between these emotions. Cross-cultural recognition studies provide a useful indication of the extent to which cross-cultural possibly psycho-physiologically influenced vocal cues of emotion exist, particularly with regard to basic emotions. However, they provide no evidence as to what these cues may be, or to what extent the acoustic cues of different emotions are similar across different cultures.

So, whilst emotion is a key element of indexical information signalled in the voice, there remains a lack of cross-cultural research in this area to ascertain the extent to which the same vocal cues can be decoded as the same emotion by different cultures, and how far the same cues may be interpreted differently. Considerable further cross-cultural research is needed to compare in-group and cross-cultural decoding rates in the

search for vocal cues which may be pan-cultural and those which are culturally influenced.

> …more comparative and cross-cultural research will be necessary before definitive claims can be made about how linguistic and/or cultural similarity influence emotional communication in the voice or through other channels… (Pell et al., 2009b, p.433)

There is then, a lack of research comparing vocal expression and recognition of emotion across different cultures and studies which include acoustic analysis are particularly sparse. In particular, considerable further research is required comparing vocal cues of emotion in unrelated cultures, if we are to shed more light on the existence of pan-cultural characteristics of vocal correlates of emotion.

In cross-cultural studies where acoustic analysis has been conducted, data from only one culture has tended to be analysed. As Scherer et al. (2001, p.89) commented, where "only encoders from one single culture (country and language)" are included, one "cannot draw any conclusions with respect to the universality of vocal emotion encoding" and suggested that acoustic analysis of data obtained from balanced studies, including both encoding and decoding experiments by subjects from more than one culture, will help to draw stronger conclusions regarding pan-cultural characteristics of vocal profiles of emotion.

Early studies such as Skinner (1935), and Scherer, Wallbott, Tolkmitt and Bergmann (1985) included acoustic analysis and others (Eldred & Price, 1958; Davitz, 1964) included auditory analysis of data. Van Bezooijen (1984) suggested that for acoustic parameters of emotion to be considered "meaningful and appropriate for use as standardised tools for describing emotional expressions" (p.57), correlations should exist between perceptual and acoustic parameters. This is about looking at the proximal cues element of Scherer's adaptation of The Brunswikian Lens Model (Scherer, 2003), discussed in Chapter One. Van Bezooijen (1984) is a rare example of a study which incorporates both decoding tests and acoustic and auditory analysis. More recently, where vocal cues have been analysed, this has tended to be through instrumental rather than auditory analysis, probably largely due to the considerable advances in the area of instrumental phonetics particularly over the past twenty years.

Juslin and Laukka (2003) found only twelve cross-cultural studies in their meta-analysis, and only seven of these included acoustic or auditory analysis (Abelin and Allwood, 2000; Breitenstein et al., 2001; Chung, 2000; Juslin and Laukka, 2001; Nakamichi et al 2002; Scherer et al, 2001; Van Bezooijen, 1984). There have been some more recent cross-cultural studies in this area which have included acoustic analysis, although evidence of cross-cultural similarities and differences remains sparse. Examples include Erickson, 2006, 2010; Erickson et al, 2008a; Erickson et al, 2008b; Laukka et al. 2010; Pell et al (2009a); Pell et al (2009b); Thompson and Balkwill (2006); Occasionally, studies have used manipulated, synthesised data to test the influence of different parameters on the emotions decoded (Breitenstein et al., 2001; Burkhardt et al. 2006).

Inconsistencies in the testing methods employed across different cultures may differentially influence the acoustic characteristics of the data gathered. Cross-cultural studies tend to be particularly challenging in that they have additional methodological issues involving trade-off between consistency and authenticity of the acoustic data to be analysed. It should be considered that using a method for gathering data which facilitates greater cross-cultural consistency may also reduce the authenticity of the data, since this tends to be gathered from more controlled experiments, rather than from 'ecologically valid' settings. To date, it is usually the case that cross-cultural studies employ more controlled testing procedures, given the need for cross-cultural consistency to reduce the likelihood of inconsistencies influencing the analysis of acoustic cues of emotion in the data. However, the few balanced cross-cultural studies which have been conducted tend not to include acoustic analysis of the data and do not employ the same method for gathering data from both cultures. The data tends to be phonetically and phonotactically inconsistent, which may also influence results of acoustic analysis. Issues surrounding methods of data collection and their implications for the trade-off between consistency and authenticity are discussed further in Chapter Three.

2.2 Vocal cues of basic emotions

Most previous studies, both mono-cultural and cross-cultural, have tended to focus on measurement of f0, intensity and/or duration and have included a selection of the following parameters: f0 mean (Hz), f0 range (Hz), and f0 variability (Hz), intensity

mean (dB) and intensity range (dB) and measurements of duration, usually speech rate (syllables per second). Pell et al. (2009b) analysed f0 mean, f0 range and speech rate. The cross-cultural study by Braun and Oba (2007) focussed specifically on comparing the use of duration as a vocal cue to emotion. They included analysis of speech rate (syllables per second), articulation rate (syllables per second excluding pauses) and various measures relating to pauses.

Despite the fact that the evidence for stress-timing is contested territory, stress timing versus syllable timing is often referred to and is regarded as a tendency rather than a hard and fast rule. With this caveat in mind, English tends to be regarded as a stress-timed language (Pike, 1946; Abercrombie, 1967), in which there is a tendency to isochrony between stressed syllables which influences syllable length. Japanese has sometimes been referred to as a syllable-timed language (Catford, 1977; Ohata, 2004), in which syllables tend to be of similar length, leading to different rhythmic structure from that found in syllable-timed languages. Japanese is often described as a mora-timed language (Ladefoged, 1975; Otake, Hatano, Cutler & Mehler, 1993), where there is a tendency towards isochrony between one mora and another. It has also been argued (Warner & Arai, 2001) that there is regional variation in whether a Japanese accent is mora-timed or syllable-timed. Braun and Oba (2007) refer to Japanese as a syllable-timed language, commenting that the term "mora" is still under discussion.

Unlike English, Japanese has been described as a pitch accent language (Cutler & Otake, 1999), where pitch accents vary depending upon whether or not a word is stressed and on the number of morae in a word. Pitch accents have been found to distinguish word meaning and possibly grammar and to vary depending upon sociological factors such as age, social group, perception and reaction to social stigma, as well as geographical region.

Some studies have included other acoustic parameters, such as voice quality, f0 contour, intensity variability, high frequency energy and voice quality, although these have tended to be mono-cultural studies. There is as yet little consensus as to the importance of these variables in the distinction of specific emotions, although more evidence has been found for the possible role of voice quality and f0 contour in the communication of emotion. Mozziconacci (1998; 2002) reported the influence of f0 contour on the vocal expression of emotion. The mono-cultural study by Banziger and Scherer (2005, p.265) found tentative evidence of the influence of "relative height of local f0 excursions,

contour "shape" and final fall" in the vocal expression of emotion. See Auberge, Audibert and Rilliard (2004), Mozziconacci (1998), Mozziconacci (2002) and Yanushevskaya, Gobl and Ní Chasaide (2005) for further discussion of the possible role of f0 contour in the vocal expression of emotion.

In relation to this, Fonagy (1981) suggested that basic emotions may be distinguished by prosodic differences, but that subsets of emotions can be distinguished by different voice qualities. Voice qualities are difficult to define objectively and to measure precisely. The development of the Vocal Profile Analysis Scheme (VPAS) by Laver et al. (1981) made a significant contribution to the conceptualisation and systematisation of voice quality. Nevertheless, there is still considerable research to be done as to how to measure what we perceive as different voice qualities instrumentally.

There has been controversy as to the relative roles of prosodic and phonatory features in the vocal communication of emotion. Scherer (1986) commented,

> although fundamental frequency parameters.....are undoubtedly important in the vocal expression of emotion, the key to vocal differentiation of discrete emotions seems to be voice quality,…, acoustically determined by the pattern of energy distribution in the spectrum. (p.145)

Six years later, Murray and Arnott (1993) made the following comment based on their research:

> It seems that pitch envelope (i.e. the level, range. shape and timing of the pitch contour) is the most important parameter in differentiating between the basic emotions, and it is the voice quality which is important in differentiating between the secondary emotions. (p.1106)

Gobl and Chasaide (2003) reviewed previous literature on the possible association between voice quality and emotion and found that whilst they found some evidence that both Angry and Joy may be vocalised with a tense voice, as had been suggested by Scherer (1986), voice quality tended to be more associated with milder emotions than stronger, basic emotions. As explained above, the term 'Joy' rather than 'Happy' is sometimes used as is the case in this study and in several other studies as indicated below. Recall that Joy is sometimes used as a synonym for Happy. See Gobl and Chasaide (2003) and Yanushevskaya et al. (2005) for further discussion of the possible role of voice quality in the vocal communication of emotion. Since this thesis is concerned with basic emotions, there is no further discussion of voice quality here.

The few studies made of affect bursts also suggest that the acoustic profiles of affect bursts for basic emotions may have similarities to those found for verbal and pseudo-utterance vocalisations (for further information see Scherer, 1994; Schröder, 2000; 2003; Schröder et al., 2006; Sauter & Scott, 2007; Sauter et al., 2009).

Research to date suggests that characteristics of f0 have been found to be the main features distinguishing emotions when they are expressed vocally as shown below. However, problems, inconsistencies and lack of explicitness and consensus across studies in methods of measurement of f0 make it difficult to compare results, in addition to potentially influencing results of acoustic analysis and cross-cultural comparison. This is an area which has previously gained little attention. Increased consistency would facilitate direct comparison between studies and possibly provide more evidence of significant distinctions between the vocal profiles of different emotions.

Methods for measuring f0 do not tend to be detailed in previous studies in the area of emotion in speech. Indeed, it is often the case that it is not explicitly stated whether whole track or percentile measurements are taken and if percentile, which percentile measurements were included. Thomson and Balkwill (2006) and Anolli et al. (2008) do state that they calculated f0 mean and f0 range measurements from the whole f0 track whilst Banse and Scherer (1996) for example, state that they used 25^{th} and 75^{th} percentile f0 measurements. Mennen et al. (2012) is not a study of the vocal correlates of emotion; however, it is a cross-cultural study comparing differences in f0 range according to different measures in English and German. The study compares various measures of f0 range, including whole track (maximum minus minimum f0) and percentile measures. Although whole track measures ($p=0.000$), 90 percentile ($p=0.008$), and 80 percentile ($p=0.139$) range f0 measures were all significant in showing differences in f0 range between female English and German speakers, the effect size was greatest for whole track measures ($r=471$).

Investigations specifically of the vocal correlates of emotion, including, for example, Pell et al. (2009b) and the considerable body of work by Scherer, generally use Hertz as the unit of measurement of f0. Using a common unit of measurement helps to make results more directly comparable across different studies. There is to date no cross-cultural study of the vocal expression of emotion, of which the author is aware, which has investigated the relative merits of different measures of f0 in this area.

f0 tracking errors have occasionally been considered in the few recent studies which have included investigation of f0 contour, such as Banziger and Scherer (2005) and Yanushevskaya, Gobl and Ní Chasaide (2005). Pell et al. (2009b) commented that they manually inspected the whole f0 track and corrected any "obvious" 'doubling or 'halving' errors in the track before calculating f0 measurements. The authors found they needed to correct "approximately .05% of all tokens" (Pell et al., 2009b, p.422).

The issue of correction of f0 extraction errors (sometimes referred to as 'f0 track extraction errors') and how to correct for these is an area of research in itself (Batliner et al., 2007; Steidl et al., 2008; Owren & Bachorowski, 2007) and a topic for considerable debate between researchers on blogs such as the praat-users group. These errors will tend to occur whichever software is used to extract f0 and usually relate specifically to jumps, especially octave jumps, in the f0 track; sections auditorily rated as unvoiced which are tracked as voiced and sections auditorily rated as voiced which are not tracked. As Steidl et al. (2008) comment, "Since the manual correction is geared to human perception, a better term instead of 'correction' would be 'smoothed' and adjusted to human perception" (Steidl et al., 2008, p.529).

Whether and how such 'errors' are corrected can affect f0 measures such as f0 mean, f0 maximum, f0 minimum and f0 range which are commonly extracted in acoustic analysis within studies of emotion. Again, lack of consensus across studies can cause difficulties when attempting to compare results. Differences in results can be compounded by differences in how measurements are taken. Being explicit about whether pitch extraction 'errors' have been 'corrected' and if so, how, will give greater consistency, facilitating more direct comparison between studies and possibly more evidence of significant distinctions between the vocal profiles of different emotions.

Banziger and Scherer (2005) indicated that they made no correction where the algorithm did not detect voicing and sections with 'anomalous' periodicity were manually unvoiced within the 'Pitch edit…' option.

According to Boersma (2004), the automatic 'kill octave jumps' command "flattens out all the large pitch jumps, even the real ones" and he advised that smoothing "does not select a correct frequency from a number of candidates. Like "Kill octave jumps", it also rarely improves the pitch contour." Boersma (2005) continued that the smoothing option "low-pass filters the pitch curve. It cannot be used to filter away erratic values, as you will have noticed by trying." Banziger and Scherer (2005) made auditory

judgements and manually corrected octave jump ‘errors’ in the pitch contour. Scheffers (1988) commented that low-pass filtering does not remove from pitch contours irregularities which do not have any relation to the perceived pitch track and that low-pass filtering also affects "the slope and onset and offset moments of the important movements" (1988, p.981). Pfitzinger et al. (2009) claimed that Momel (modelling melody) produces 'a smoothed version that is perceptually indistinguishable from the original and supposedly void of microprosodic fluctuations' (2009, p.2455).

A further variable is how acoustic parameters are measured. There are various possible units of measurement for f0 and each has its problems and advantages. Some are more geared to production of vibration of the vocal chords (Hertz) and others how variation in f0 is perceived (Semitones). Hertz is about periodicity in the waveform and semitones relate more to a perceptual dimension.

In this area of research, f0 is measured in Hertz as investigators are searching for the phonetic correlates of emotional states f0 is also generally measured in Hertz in the few cross-cultural studies which have included measurement of f0, for example, in Pell et al. (2009b) and in the considerable body of work by Klaus Scherer. Furthermore, a recent study by Mennen et al. (2008; 2012) investigated whether using different scales of measurement for f0 made a difference to correlations with listener judgements. Whilst not concerned with researching specifically the vocal correlates of emotion, according to Mennen et al. (2008):

> Finally, we investigated whether correlations with listener judgements differed for three scales of measurement: linear Hertz, logarithmic musical semitones and the psychoacoustic ERB scale. Contrary to expectation, the unit of measure did not appear to play a clear role in the listener judgements. For the English listeners, we found no difference between the three scales of measurement. However, the ERB and Hertz scale appeared to be better units for measuring span than the musical semitone scale for the German listeners… (pp.18-19).

Whilst using a common unit of measurement helps to make results more directly comparable across different studies, other issues remain, including, for example, issues in measurement of f0, length of utterance, phonetic and phonotactic differences, normalisation for gender differences and many other issues related to encoding methods, which are discussed in Chapter Three.

As discussed, a further complicating issue when attempting to compare results of acoustic analysis across different studies has been the lack of cross-cultural phonetic and phonotactic consistency in the data. This will be discussed further in Chapter Three.

Where previous studies have included acoustic analysis of encoded data, these studies have tended to investigate the acoustic cues for one or more of four basic emotions, focussing on Happy, Sad, Angry and Fearful. 'Joy' has sometimes been used to refer to a more intense form of 'Happy'. However, where 'Joy' is included, this is usually employed as a synonym for 'Happy'. Evidence of distal acoustic cues from previous studies has been mainly from mono-cultural studies and tends to highlight mainly the basic distinction between cues for high and low arousal which is evident in Table 2.1 below, which is a summary of Table 7 from the meta-analysis by Juslin and Laukka (2003, pp.792-5). The study was a meta-analysis and as such, was an attempt to find and report common ground in results of studies up to 2003, the vast majority of which were mono-cultural. The table indicates general patterns within which there is a considerable amount of noise. Evidence, including some cross-cultural evidence since 2003, has provided further evidence, which is discussed in the context of each emotion below. Results are included in this adapted table only if evidence was found in at least five studies.

There is preliminary evidence in cross-cultural studies of pan-cultural tendencies in the use of certain acoustic parameters, in particular f0, in the encoding of what are regarded as basic emotions. Most of the evidence on acoustic cues of emotion emerges from studies which perform acoustic analysis of the vocal expression of emotion by single, usually western, cultures. However, a mono-cultural study of Japanese vocal expression of emotion (Hayashi, 1999), for example, found that f0 had an important role in communicating emotions in Japanese. Some evidence of cross-cultural differences in how these cues are used to express emotion has also been reported.

However, whilst Table 2.1 shows a clear distinction in the acoustic cues which have been found to distinguish between high arousal (Angry, Fearful, Happy) and low arousal basic emotions (Sad), according to the meta-analysis by Juslin and Laukka (2003), these cues did not distinguish between Angry and Happy, which are both high arousal emotions. They found that Angry and Happy were only distinguished from Fearful, the other high arousal emotion analysed, by the narrow f0 range used for Fearful, compared to the wide f0 range employed for Happy and Angry. It should be noted, however, that results for f0 range for Fearful vary, possibly due to different types of Fearful being vocalised. There is evidence that whilst f0 range appears to be narrow for mild fear (e.g. anxiety), strong fear (e.g. panic) is vocalised with a wider f0 range, in common with other high arousal emotions. It appears that the two-way distinction for

each acoustic parameter (high/low, wide/narrow, fast/slow, up/down) is insufficient to distinguish between the high arousal emotions. In addition, the fact that only one low arousal emotion has tended to be included in previous studies means that there is a lack of evidence on the extent to which these acoustic cues vary between one low arousal emotion and another.

Emotion	**Acoustic cue**	**Category**
Angry	Speech rate	Fast
	Mean intensity	High
	Intensity range [1]	Wide
	High frequency energy	High
	Mean f0	High
	f0 range [2]	Wide
	f0 contours[3]	Up
Fearful	Speech rate	Fast
	Mean intensity	High
	Intensity range [1]	Wide
	High frequency energy	High
	Mean f0	High
	f0 range [2]	Narrow
	f0 contours[3]	Up
Happy	Speech rate	Fast
	Mean intensity	High
	Intensity range [1]	Wide
	High frequency energy	High
	Mean f0	High
	f0 range [2]	Wide
	f0 contours[3]	Up
Sad	Speech rate	Slow
	Mean intensity	Low
	Intensity range [1]	Narrow
	High frequency energy	Low
	Mean f0	Low
	f0 range [2]	Narrow
	f0 contours[3]	Down

Table 2.1: Acoustic parameters of basic emotions: evidence from mainly mono-cultural studies. Adapted from Table 7 in Juslin & Laukka (2003, pp.792-5)

1. The term "intensity variability" is used by Juslin & Laukka (2003, p.792) to refer to intensity range.
2. The term "f0 variability" is used by Juslin & Laukka (2003, p. 792) to refer to f0 range.
3. f0 contour, sometimes referred to as 'pitch track' is a graph of fundamental frequency plotted against time.

The acoustic distinction between high and low arousal emotions can help to explain patterns of confusion in decoding experiments, since errors often occur along the dimensional lines of arousal. Happy, Angry and Fearful, for example tend to be confused with each other more than they are confused with Sad. Low arousal Sad tends to be vocalised with a slow speech rate, low mean f0, narrow f0 range, little f0 variability, low mean intensity and a narrow range of intensity, whilst high arousal Happy, Angry and Fearful tend to show the opposite patterns, although there is evidence of a narrow f0 range for Fearful. The mono-cultural study by Banziger and Scherer (2005) reported that f0 mean and f0 range were sufficient to account for level of arousal in emotion portrayal. One feature of much previous work in this area, which has been commented upon by Juslin and Laukka (2003), is the lack of standardisation in encoding and decoding procedures and reporting across different studies. They also found in their meta-analysis of mono-cultural and cross-cultural studies of the vocal communication of emotion that the frequent absence of tests for statistical significance meant that results were difficult to compare, although distinctions between the acoustic cues used to signal high and low arousal emotions appeared clear.

Thomson and Balkwill (2006) reported that measures of mean and range of f0, mean and range of intensity and speech rate varied as a function of intended emotion in the cultures they investigated (Tagalog, English, Chinese, Japanese and German**).** However, the only significant variation due to culture was the association between range of intensity and emotion. (Pell et al. (2009b) investigated in-group decoding rates and acoustic cues for Happy, Sad, Angry and Fearful for German, Hindi, English and Arabic using the same encoding procedure for each culture, making results comparable. It should be noted that no cross-cultural decoding tests were conducted so cross-cultural perception of emotion vocalisations was not tested. However, they concluded that f0 mean, f0 range and speech rate distinguished well between Happy, Sad and Fearful in German, Hindi, English and Arabic, although these parameters were less predictive of Angry. They also suggested that the inclusion of other acoustic parameters such as intensity may help to distinguish Angry vocalisations. More detailed information on their findings for each emotion is given in the subsections on each emotion below. The study by Pell et al. (2009b) revealed cross-cultural similarities, particularly in the use of f0, to communicate emotion, providing evidence supporting the existence of universal and possibly phylogenetic tendencies, commenting “These modal tendencies likely reflect properties of natural, coded signals which are used to communicate emotions in

speech (Wilson & Wharton, 2006) and which are shared in large measure across languages" (Pell et al., 2009b, p.434).

Pell et al. (2009b) also found cultural differences in the use f0 and duration to communicate emotion. Preliminary evidence found in previous studies of cross-cultural similarities and differences in the vocal cues used to express emotion is discussed below.

As is evident from Table 2.1, whilst some common tendencies have emerged in what appear to be the acoustic cues of basic emotions, particularly in relation to the distinction between high and low arousal emotions, considerable further research is required in the search for vocal cues of emotion and the extent of cross-cultural similarity and difference.

It has been suggested that East Asian cultures show more restraint in the vocal expression of emotion than other cultures (Anolli et al., 2008; Mesquita & Walker, 2003; Soto, Levenson & Ebling, 2005). Mesquita and Walker (2003) suggest that East Asian cultures focus on relational harmony and the importance of taking one's proper place and that Western cultures focus more on the individual. Evidence in Braun and Oba (2007) suggests that overall the two Japanese speakers they included in their study produced a faster speech rate and articulation rate than did the American English and German speakers for all emotions. Evidence of cross-cultural similarities and differences in the vocal cues of emotion remains sparse. However, examples of findings to date are presented below, including evidence demonstrating possible emerging trends and other apparently contradictory evidence. Possible vocal correlates of Calm are suggested based on the dimensional similarities with Happy and Sad.

2.2.1 *Cross-cultural similarities and differences in the vocal cues of Happy*

Previous studies, including the few cross-cultural studies which have conducted acoustic analysis have generally found that the acoustic characteristics of Happy are distinct from those of Sad, which is a low arousal emotion.

According to the meta-analysis of mainly mono-cultural research by Juslin and Laukka (2003), Happy, like Angry and Fearful, are vocalised with a fast speech rate, high mean intensity, wide intensity range, high high frequency energy, high mean f0 and rising f0

contour. In addition, f0 range may distinguish Happy from Fearful since there is evidence that Happy, like Angry, has a wide f0 range whilst Fearful has a narrow f0 range. Gobl and Chasaide (2003) found evidence of an association between Joy and a tense voice quality, although the strongest association with tense voice was with Angry. As discussed in 3.9.4, whilst Happy is the most reliably recognised emotion cross-culturally from facial expression (Ekman, Sorenson, & Friesen, 1969; Izard, 1994), there is some evidence that Happy is more difficult to recognise cross-culturally from the voice than other basic emotions, Angry and Sad being the most reliably recognised emotions from vocal cues alone.

Thompson and Balkwill (2006) found that cross-culturally, Joy was vocalised with a wider f0 range than other emotions. Mean intensity was also higher for Joy and Anger than for Fearful and Sad. For each acoustic parameter the authors reported that there was no significant cross-cultural difference in the association between emotion and acoustic parameter except for the relationship between range of intensity and emotion, which did vary significantly from one culture to another (2006).

Anolli et al. (2008) reported that the Italian and Chinese vocal profiles of Joy in their study were significantly different. Compared to the Chinese data, the Italian data was characterised by medium pauses, slower speech rate, higher mean f0, higher standard deviation of f0 and higher mean intensity. No definition is given by the authors as to what they mean by the term 'medium pauses' or what the 'baseline' may be. Their Chinese data for Joy demonstrated short pauses, medium mean f0, medium standard deviation of f0 and medium mean intensity. Native Chinese speakers also had a faster speech rate, and articulation rates as well as greater variations in voice intensity.

Braun and Oba (2007) analysed several measures of duration for Japanese, German and American English. Their study included both native and non-native decoding tests which tested reliability of the data. They pointed out cross-cultural differences in the use of different measurements for timing to communicate Joy, Sad, Angry and Fearful. They reported that Japanese vocal expressions of emotion tended to be faster than American English and German expressions. Pell et al. (2009b) found that Arabic vocalised Happy with the slowest speech rate compared to the other emotions they investigated (Sad, Angry and Fearful), which is unusual since speech rate is usually found to be slower for Sad, which is a low arousal emotion.

Pell et al. (2009b) found evidence that the acoustic profile for Happy varied cross-culturally. English Happy tended to be vocalised with a mid f0 mean, which was significantly distinguishable from Anger, Fear and Sad. There was also a tendency for English Happy to be vocalised with a mid level f0 range. The mid speech rate of English Happy was significantly slower than Fear but significantly faster than Angry and Sad.

In contrast, in the German data, Happy was vocalised with a higher f0 mean than Fear and a significantly higher f0 mean than Sad. German speakers tended to use a wider f0 range for Happy than for Anger, Fear and Sad. Happy, along with Angry demonstrated a significantly faster speech rate than Fear and Sad, which did not differ.

In Hindi, on the other hand, Happy tended to be vocalised with a mid f0 mean, significantly distinguishable from Anger, Fear and Sad, significantly wider f0 range than Fear, a significantly slower speech rate than Anger and Fear and a significantly faster speech rate than Sad.

In their Arabic data, Happy was vocalised with a significantly lower f0 mean than Fear and a significantly higher f0 mean than Anger and Sad. Happy also showed a significantly wider f0 range than Sad, also demonstrating a wider f0 range than Anger. Arabic vocalised Happy with a significantly slower speech rate than Fear, Anger and Sad. This is unexpected as the slowest speech rate has tended to be found for sad in previous studies of other languages.

2.2.2 *Cross-cultural similarities and differences in vocal cues of Sad*

Previous studies have found that low arousal Sad tends to demonstrate a pattern of acoustic parameters which is fairly distinct from other high arousal basic emotions across all cultures studied. According to the meta-analysis by Juslin and Laukka (2003), Sad is vocalised with a slow speech rate, low mean intensity, narrow intensity range, low high frequency energy, low mean f0 and falling f0 contour, Happy, Angry and Sad showing the opposing values. Sad and Fearful have a narrow f0 range, whilst Happy and Angry have a wide f0 range. Sad, Angry and Fearful have microstructural irregularity, whilst Happy shows microstructural regularity, although there is as yet little evidence to support this.

Braun and Oba (2007), for example, analysed several measures of duration of Joyful, Sad, Angry, Fearful and Neutral vocal expression in American English, German and Japanese. Whilst there were no tests reporting statistical significance, descriptively, Sad had the slowest speech rate for all of the cultures in the study. As discussed in 3.9.4, Sad tends to be recognised with a high level of accuracy from vocal cues alone, both in-group and cross-culturally.

Sad tends to be consistently distinguished cross-culturally from other basic emotions by the acoustic characteristics described above. However, occasional exceptions have been reported. For example, Pell et al. (2009b) found that for English, German, Hindi and Arabic, Sad was vocalised with a lower f0 mean, narrower f0 range and slower speech rate than Happy, Angry and Fearful, except for Arabic, where participants vocalised Happy with a significantly slower speech rate than other basic emotions including Sad. Thompson and Balkwill (2006) also found that German participants vocalised Sad with a greater "event density" (number of waveform peaks divided by utterance duration), in other words a faster speech rate, than Joy. Chinese participants in this study also vocalised Sad with a faster speech rate than both Joy and Fearful.

2.2.3 *Cross-cultural similarities and differences in vocal cues of Angry*

According to the meta-analysis by Juslin and Laukka (2003), Angry, Happy and Fearful are vocalised with a fast speech rate, high mean intensity, wide intensity range, high high frequency energy, high mean f0 and rising f0 contour. Angry and Happy have a wide f0 range whilst Fearful has a narrow f0 range. Angry and Fearful have microstructural irregularity, whilst Happy shows microstructural regularity, although there is as yet little evidence to support this. Previous studies have generally found that the acoustic characteristics of Angry are distinct from those of Sad.

As discussed in 3.9.4, Angry tends to be recognised with a high level of accuracy from vocal cues alone, both in-group and cross-culturally. However, evidence of acoustic cues of Angry which distinguish this emotion from other high arousal emotions has proved elusive. Thompson and Balkwill (2006) found that Angry had a higher mean pitch and mean intensity than Fear and Sad and a faster speech rate than Sad for all of the cultures they studied (Japanese, Tagalog, German, English and Chinese). Similarly,

Braun and Oba (2007) found evidence of Angry having a faster speech rate for the cultures they investigated (Japanese, American English and German).

Pell et al. (2009b) found that the acoustic pattern for Angry was culture-dependent across the cultures they studied. For example, f0 mean was relatively high in Hindi, mid in English and German, and low in Arabic and f0 range was relatively wide in Hindi, mid in English and German and narrow in Arabic. The authors suggested that some of these differences may be due to variation in the type of Anger expressed. They argued that "hot anger", characterised as being high arousal, would tend to have high f0 mean and high intensity and "cold anger", having a low level of arousal, would tend to be vocalised with a mid or low f0 mean and a mid or narrow f0 range. On the other hand, the authors also point out that the different profiles may demonstrate cross-cultural differences.

Whilst Angry was recognised reliably from voice only cues in their decoding test (Pell et al., 2009b), the acoustic parameters they analysed (speech rate, f0 mean and f0 range) did not differentiate Angry well from the other emotions (Sad, Happy, Fear) for English, German or Arabic, although these parameters did distinguish Angry in their Hindi examples. One possible explanation they give is that the different acoustic patterns in the Hindi examples may have been due to the expression of 'hot anger' compared to the expression of 'cold anger' by the other cultural groups, arguing that 'hot anger' may be more distinguishable from the other emotions than 'cold anger'.

However, they also suggest that the inclusion of other acoustic variables such as intensity variation, voice quality and energy distribution may improve the acoustic classification of Angry expressions since previous studies (Banse & Scherer, 1996; Fónagy & Magdics, 1963; Gobl & Chasaide, 2003; Sobin & Alpert, 1999; Williams & Stevens, 1972) have found these parameters to be relevant in the vocal expression of Angry.

Anolli et al. (2008) found no statistically significant differences between Chinese and Italian participants in acoustic cues of Angry. However, they commented on a purely descriptive basis, that speech rate, mean f0, f0 variability and mean intensity, all had higher values for Italian than for their Chinese participants.

2.2.4 *Cross-cultural similarities and differences in vocal cues of Fearful*

According to the meta-analysis by Juslin and Laukka (2003), evidence of the acoustic characteristics of Fearful is similar to that for other high arousal emotions. Fearful, Angry and Happy are vocalised with a fast speech rate, high mean intensity, wide intensity range, high high frequency energy, high mean f0 and rising f0 contour. Fearful has a narrow f0 range, whilst Angry and Happy have a wide f0 range. Previous studies have generally found that the acoustic characteristics of Fearful are distinct from those of Sad, except that both Fearful and Sad have a narrow f0 range.

Contradictory results have been found for vocal features of utterances perceived as Fearful in previous studies. It has been argued (Banse & Scherer, 1996; Juslin & Laukka, 2001; Juslin & Laukka, 2003; Pell et al., 2009b) that conflicting acoustic profiles for Fearful could be explained by different types of Fearful being vocalised, some representing anguish or anxiety and some representing 'panic fear'.

Pell et al. (2009b) found that all the cultures they studied (English, German, Arabic and Hindi) showed a relatively high f0 mean for Fearful. Anolli et al. (2008) also found evidence that both Italian and Chinese participants vocalised Fearful with a high mean f0.

Whilst Pell et al. (2009b) found that all the cultures they studied showed a relatively high f0 mean for Fearful, they reported that the German group in their study produced Fearful with a slow speech rate (syllables per second) and 'a markedly narrow f0Range' (Pell et al. 2009b, p.432), compared to Happy, Sad and Angry, whilst the English, Hindi and Arabic groups all used a fast speech rate and a wider range. The authors suggested that the German speakers may have been expressing 'panic fear' whilst the English, Hindi and Arabic speakers may have been expressing 'anguish' or 'sustained fear. They also commented that the longer length of utterances produced by the German speakers may have made the expression of 'panic fear' feel 'unnatural' and suggest that

> Controlling better for utterance complexity/ length and for the… intensity of emotional expressions under study remain an ongoing challenge for researchers, although these factors are likely to influence the acoustic structure of emotional speech in a significant manner (Pell et al., 2009b, p.432).

Although Anolli et al. (2008) found evidence that both Italian and Chinese participants vocalised Fearful with a high mean f0, the Italian participants also used a fast speech

rate, medium f0 standard deviation and high mean intensity whilst their native Chinese participants vocalised Fearful with medium speech rate, low f0 standard deviation and medium mean intensity.

2.2.5 *Vocal cues of Calm*

Until recently previous studies of the vocal communication of emotion have not included Calm as a category, studies tending to include only one emotion which is classed as low arousal (Sad) and one classed as positive valence (Happy).

A recent monocultural study by Livingstone et al. (2013) includes Calm, alongside Happy, Sad, Angry and Fearful. This is a monocultural study comparing acoustic features of emotion in speech and song. They found that in common with Sad, Calm was vocalised with a slower speech rate, narrower f0 range and lower mean intensity than the high arousal emotions (Happy, Angry and Fearful). However, they do not appear to have found evidence of any acoustic cues shared by the two positive valence emotions, Calm and Happy. Further research is required before any stronger conclusions can be drawn on this, including cross-cultural research, which may highlight universal tendencies or cultural differences in the acoustic properties of Calm.

2.3 Summary

Whilst there is some evidence from cross-cultural recognition that subjects from different cultures can recognise distinct basic emotions with a high degree of accuracy, considerable further research is required into the acoustic parameters which may cue these distinctions. In particular, if we are to further knowledge of possibly innate, pan-cultural tendencies, more investigation of cues used by speakers from unrelated cultures is required.

Taking into account that mono-cultural studies have found that the major prosodic cues of duration, pitch and loudness in the form of instrumental measures of speech rate, f0 and intensity have an important role to play in the vocal communication of emotion, the investigation of these cues has been a logical starting point for some of the early cross-cultural research.

As discussed, the main distinction drawn by the acoustic parameters of f0, intensity and duration has been between high and low arousal emotions. Cross-cultural evidence has been found of f0, intensity and duration measurements tending to distinguish Sad from the other basic emotions. It is likely that this is because it is the only low arousal emotion in the set of basic emotions investigated. Interesting anomalies emerge regarding the three high arousal basic emotions, Happy, Angry and Fear. Whilst Angry tends to have a particularly high recognition rate, along with Sad, determining the vocal cues of Angry has proved elusive. Different Angry vocalisations sometimes present opposing acoustic correlates, possibly influenced by whether 'hot anger' or 'cold anger' is expressed. Apparently conflicting results have been found for the vocal cues of Fearful in previous studies. It has also been argued (Banse & Scherer, 1996; Juslin & Laukka, 2001; Juslin & Laukka 2003; Pell et al., 2009b) that this could be explained by different types of Fearful being vocalised, some expressing anguish or anxiety and others expressing 'panic'. Some previous cross-cultural studies have found that Happy has relatively low levels of cross-cultural decoding reliability compared to Angry and Sad and there is preliminary evidence of cultural variation in vocal cues signalling Happy as discussed in 2.2.1.

Cultural groups represented by western languages may have similarities in the acoustic features they use to express emotion which are not found in non-western languages. Further research which includes non-western cultures may give more insight into possible pan-cultural vocal cues of emotion. Individual differences in either decoding results or in acoustic analysis have not been considered in previous cross-cultural studies in this area. It is still very early days for cross-cultural research in this area and sample sizes have also tended to be small. This remains an area requiring investigation within the field of the vocal expression of emotion.

Constructing a balanced methodology for cross-cultural studies in this area and gathering data for decoding experiments and acoustic analysis, which is also cross-culturally symmetrical in that it is as far as possible more linguistically, phonetically and phonotactically consistent is a methodological challenge which has not yet been addressed. It is inevitable that there is a trade-off between consistency and authenticity in the method used to gather data. Controlling for consistency of data gathered for the recognition test and for acoustic analysis will tend to mean that the data will not be from natural, spontaneous speech, but from more controlled experiments. It is argued (Scherer, 2013b) that this is particularly justified in the context of current cross-cultural

studies in this area, at least until greater understanding of cross cultural similarities and differences is gained for more cultures. Further discussion of issues of consistency and authenticity is given in Chapter Three. The methodological difficulties which have led to the gaps in research and evidence in this area are detailed in the following chapter, issue by issue.

Previous cross-cultural studies have usually tended to be recognition studies, where decoding of encoded data is tested. However, the data has tended not to be analysed acoustically to ascertain the vocal correlates of the emotions recognised.

This chapter has focussed on the production data, reviewing previous research showing evidence of cross-cultural and culture-specific vocal cues emotion. If we are to study vocal correlates in the cross-cultural dimension, we need to consider how that encoding material comes to be and the perception of this material by people from different cultures. Before entering further into the study of the vocal correlates used to communicate emotion, explanation is required as to how the data was accessed for the two cultures. The many methodological issues involved in this area are quite severe and this explains the lack of cross-cultural research in this area. A major aim of this study is to address these methodological issues in order to gather data for acoustic analysis which is both encoded and decoded by each culture and which is more phonetically and phonotactically consistent than any previous study in this area. This is vital in order to further knowledge of cross-cultural similarities and differences in the vocal cues of emotion.

For clarity, these issues are discussed in detail issue by issue in Chapter Three, explaining methods used in previous cross-cultural studies in this area Findings of these studies are also discussed in the next chapter as these inform the research questions for the thesis, which are detailed in Chapter Four.

Chapter Three

Encoding and decoding methods and recognition findings in previous studies

> Whilst cross-cultural research on the facial expression of emotion can directly observe facial expression, vocal cues cannot be so directly observed and isolating vocal cues from verbal cues is also a challenge in vocal expression research. Cross-cultural research poses the additional challenge of designing a balanced methodology which allows consistency in the decoding tests presented to participants from different cultures who are also generally from different native language backgrounds as this helps maximise cultural difference (Scherer, Banse, Wallbott 2001, p.88).

3.1 Introduction

The previous chapter reviewed research on the phonetic properties of the way in which different emotions are encoded. However, before we can begin to analyse the data gathered in the present study, it is necessary to explain both how the production and recognition data were gathered for the present study and the findings and questions arising from previous recognition research which, alongside previous findings from acoustic analysis, have led to the research questions posed in the present study.

Despite the growth of research in the area of emotion in speech, one area which is still under-investigated is that of cross-cultural differentiation due to the additional methodological challenges involved. In addition to the empirical observation of vocal cues of emotion and separation of verbal cues from vocal cues, there are also translation issues to deal with in cross-cultural studies and in particular the design of a 'balanced methodology' (Scherer, Banse & Wallbott, 2001, p.88) in cross-cultural studies presents considerable challenges. Indeed, a new balanced methodology for cross cultural studies in this area has been called for, for almost two decades. Cross-cultural studies have tended to include only analysis of recognition rates and patterns in decoding data. Less commonly, a few studies have analysed acoustic cues in the production data, sometimes without a recognition test. Previous studies have therefore often potentially ignored either cross-cultural differences in acoustic cues or differences in recognition of these cues. Despite the relatively recent growth in research in this area, due mainly to methodological difficulties there remains a need for balanced cross-cultural studies which allow comparison of how the data is recognised both in-group and cross-

culturally by both cultures and analysis of vocal cues which account for similarities and differences in perception of emotion in the vocal expressions of emotion by both cultures. The empirical observation of vocal cues of emotion and the separation of verbal cues from vocal cues are issues which have presented considerable methodological challenges to researchers in this field, particularly in cross-cultural studies in this area.

Since one of the main aims of the present study was to meet the challenge of constructing a new, more effective encoding and decoding method for cross-cultural studies in this area, there is discussion in the present chapter of the issues involved, in particular the more complex difficulties faced by cross-cultural studies in this area. For clarity, each issue is discussed in turn. The present study has attempted to address these issues in a more rigorous way than has been done before. Findings from previous recognition studies are then explained.

Reviews of studies of the vocal communication of emotion include those by Bachorowski and Owren (2003), Van Bezooijen (1984), Elfenbein and Ambady (2002), Erikson (2006), Graham et al (2001), Juslin and Laukka (2003) and Scherer (2003). Cowie, Douglas-Cowie, Tsapatsoulis, Votsis, Kollias, Fellenz, and Taylor (2001), Douglas-Cowie et al. (2003), Murray and Arnott (1993), Schröder (2001) and Vervedris and Kotropoulos (2003; 2006) review previous studies in relation to speech synthesis. These reviews include a small number of cross-cultural studies. In their review in 2003, Juslin and Laukka indicated that only 12% (12) of the studies they reported in the field of vocal communication of emotion were cross-cultural studies.

Since 2003, methodological issues involved in the encoding and decoding of emotion vocalisations have continued to provide a challenge to researchers in this area. Section 3.9.1 discusses levels of accuracy in the recognition of emotion vocalisations, also including cross-cultural decoding accuracy. Conflicting findings in relation to possible differences between in-group and cross-cultural recognition accuracy are discussed in Sections 3.9.2 and 3.9.3, which explain ‘In-Group Advantage’ and the ‘Cultural Proximity Hypothesis’ respectively. Section 3.9.4 explores findings on recognition rates for specific emotions. Arguments regarding the role of verbal and vocal cues are explained in Section 3.9.5. Hints in previous research regarding the possible influence of vowel quality are discussed in Section 3.9.6. There is a small amount of evidence of the influence of utterance length in vocal emotion encoding and decoding and this is

explained in Section 3.9.7. Although related, utterance length is different from the acoustic measure of duration which is discussed in Chapter Two. There is also a little evidence of the influence of participant gender on the encoding and decoding of emotion vocalisations. This is briefly discussed in Section 3.9.8. Neither cross-cultural nor mono-cultural studies of the vocal expression of emotion have tended to test the influence of encoding or decoding participant age or social group upon encoding and decoding of vocal emotion. In any case, most studies do not have large enough participant groups to enable reliable comments upon such influence to be made.

The findings and questions arising from previous research contributed to the construction of the research questions and predictions of the present study (see Chapter Four). The design of the encoding and decoding experiments in this study (Chapter Five) aimed to answer these research questions and to address the methodological issues detailed here. The purpose of this Chapter is therefore to place this study within the context of research on the vocal communication of emotion to date, particularly cross-cultural research, by providing a review of methodological issues and findings.

3.2 Language as an indicator of cultural differentiation

What we understand by the term culture is a complex issue. In cross-cultural work in the area of the vocal communication of emotion, researchers tend to avoid grappling with it by using language as a proxy. This is not to imply that language is the cause of cultural difference which may be reflected in vocal emotion communication. There is sometimes a lack of clarity and consensus regarding use of the terms cross-cultural and cross-language. Cross-cultural studies tend to use groups of participants who are not native speakers of the language(s) spoken by the other group(s) in the study with the aim of ensuring cultural difference between groups of encoders and/or between groups of decoders. This is not to suggest that there will be no cultural diversity amongst people living in the same country speaking the same native language, however, broad differences in culture are assumed between speakers of different languages.

Cultures represented by different varieties of a language, spoken in different geographical areas with diverse historical and linguistic influences have also tended to be treated as distinct. It seems reasonable to assume that native speakers of, for example, European Spanish may be regarded as having a culture which is broadly

distinct from that of native speakers of Mexican Spanish living in Mexico or native speakers of Argentine Spanish living in Argentina. In the same way, the cultures of native speakers of British English and native speakers of US English could also be regarded as broadly distinct. Scherer (2001) included decoders from nine different cultures. These cultures were represented by six different languages; US English and British English were regarded as two distinct cultures as were Swiss French and Parisian French. Laukka et al. (2010) included native speakers of US English as encoders and native speakers of five varieties of English, including US English, as decoders.

Whilst some researchers in this field implicitly draw a distinction between different varieties of a language, other researchers group varieties together as representing the same culture. This issue requires clarification and consensus. It could be argued that it is reasonable to regard different varieties of a language as representative of different cultures and it is therefore also helpful that the variety of a language investigated is clearly stated.

3.3 Variety of cultures studied

Until recently, literature on the subject of emotion vocalisation consisted mainly of short monolingual studies, and cross-cultural studies which have been undertaken have tended to involve Western languages. In 2003, Juslin and Laukka stated that "58% of studies in emotion vocalisation used English-speaking encoders" (2003, p.777). Of the cross-cultural studies undertaken, most have investigated non-Western languages, although more recently this has started to change. The study by Albas et al. (1976) was a rare early exception, including native speakers of Cree as both encoders and decoders. The comparison of prosodic cues of emotion in Western and non-Western cultures will provide stronger evidence of any quasi-universal patterns as well as highlighting cultural differences.

More recently, non-Western cultures have started to be investigated. Thompson and Balkwill (2006) compared recognition of native Canadian English emotion vocalisations by Canadian English, German, Chinese, Japanese and Tagalog listeners. Native speakers of Japanese were included as encoders and/or decoders in studies by Braun and Oba (2007), Erickson (2006), Erickson (2010) Erickson et al. (2008a),

Graham et al. (2001), Ibrakhim (2004), Ishii et al. (2003), Kramer (1964), Menezes et al. (2010), Nakamichi et al. (2002), Rilliard et al. (2009), Van Bezooijen et al. (1983), Van Bezooijen (1984) and Yanushevskaya, Chasaide and Gobl (2011). Native speakers of Mandarin Chinese were included as either encoders or decoders in studies by Anolli et al. (2008), Dang et al. (2009), Erickson et al. (2008a), Thompson and Balkwill (2006). Experiments by Sauter et al (2010) included native speakers Himba. Native speakers of Thai were included in the study by Xu and Chuenwattanapranithi (2007). Native speakers of Korean were included as encoders and/or decoders in studies by Chung, (2000), Erickson (2006) and Erickson (2010).

There has also been concern in previous research (Scherer, 2001) as to the possible influence of American English culture, particularly Hollywood films, upon other cultures, suggesting that similarities in encoded vocal patterns and accurate cross-cultural decoding may be due to language contact rather than being due to universal psycho-physiological influence. However, Graham et al. (2001) found that decoders with considerably greater exposure to the encoder's culture and with an advanced proficiency level in the non-native language did not decode the non-native vocal expressions of emotion any more accurately than those non-native decoders with little exposure and beginner's level in the non-native language.

3.4 Number of encoders/decoders

Previous studies of emotion vocalisation have tended to include very few encoders. Frequently, only one speaker from one culture produces all of the data (e.g. Abelin & Allwood, 2000; Breitenstein et al., 2001; Chung, 2000; Dromey et al., 2005; Erickson, 2006; Erickson 2010; Ibrakhim 2004). Very few studies have considerably more encoders. Table 3.1 below gives examples of the range of numbers of encoders and decoders in previous studies.

Anolli et al. (2008) does not include a decoding test but has 29 Chinese and 19 Italian encoders. Laukka et al. (2010) has 20 encoders of each variety of English under investigation (US English, Indian English, Kenyan English, Singlish and Australian English).

Study	Encoders	Decoders
Abelin and Allwood (2000) Abelin (2004)	Swedish (1) Spanish (1)	Swedish (35), English (12), Finnish (23), Spanish (23). Swedish (15)
Anolli et al. (2008)	Chinese (29), Italian (19)	No decoding test
Breitenstein et al. (2001)	German (1)	German (35), English (30)
Chung (1999), (2000)	Korean (1)	Korean (10), US Eng (10), French (10)
Dang et al. (2009)	Japanese (15)	Experiment 1. Japanese (17), Chinese (13), US Eng (15) Experiment 2. Japanese (13), Chinese (13)
Dromey et al. (2005)	English (1)	142 subjects: Varieties of English and 21 other languages
Erickson (2006) Erickson (2010)	Japanese (1) Korean (1)	20 Japanese (20); US English (20), Korean (9) Japanese (13), Korean (12), US Eng (15)
Ibrakhim (2004)	Japanese (1)	Japanese (10), Russian (8)
Laukka et al. (2010)	US Eng (20) Indian Eng (20) Kenyan Eng (20) Singlish (20) Aus Eng (20)	US Eng (12)
Nakamichi et al. (2002) Pell et al. (2009b)	Experiment 1.Brit Eng (5) Experiment 2. Japanese (5) English (4), German (4), Hindi (4), Arabic (4)	Experiment 1. Japanese (8) Experiment 2. Japanese (19); US Eng (15) No cross-cultural decoding. In-group decoding only: English (24), German (24), Hindi (20), Arabic (19)
Scherer et al. (2001)	German (4)	German (70), Swiss French (45), Brit Eng (40), Dutch (60), US Eng (32), Italian (43), French (51), Spanish (49), Indonesian (38)
Thompson and Balkwill (2006)	Canadian Eng (2), German (2), Chinese (2), Japanese (2), Tagalog (2)	Canadian Eng (20)

Table 3.1: Examples of the range of numbers of encoders and decoders in previous studies.

There has also been considerable variation across studies in the number of decoders, where a decoding test is included. For example, Chung (2000) tested 10 Korean, 10 US English and 10 French listeners, Ibrakhim (2004) included 10 Japanese and 8 Russian

decoders, Nakamichi et al. (2002) tested 8 Japanese listeners in one test and 19 Japanese and 15 US English speakers in another test, Thompson and Balkwill (2006) included 20 Canadian English speakers. Pell (2009b) included 24 English, 24 German, 20 Hindi and 19 Arabic in-group decoders. 70 German, 45 Swiss French, 40 British English, 60 Dutch, 32 American English, 43 Italian, 51 French, 49 Spanish, and 38 Indonesian native speakers were recruited as decoders by Scherer et al. (2001) to decode emotions encoded on pseudo-utterances by 4 native speakers of German.

The profile of participants in previous studies has varied in terms of gender, social class and age, although samples have usually not been large enough to consider the influence of these factors. This is discussed further in section 3.9.8.

Individual differences in either decoding results or in acoustic analysis have also not been considered in previous cross-cultural studies in this area. It is still very early days for cross-cultural research in this area and sample sizes have also tended to be small.

3.5 Balance and symmetry

A balanced study is one in which encoding tests are performed by native speakers from different cultural groups and both in-group and cross-cultural decoding tests are also performed by native speakers from these cultural groups. The term ‘symmetrical’ is used in the present study to denote a balanced study which has a consistent design cross-culturally for both the encoding test and the decoding test.

In previous studies, data has tended to be encoded by one cultural group and presented to several others for decoding or data has been encoded by several cultural groups and presented for discrimination to a single cultural group. Most cross-cultural studies of the vocal communication of emotion have tended to analyse the ability of participants from more than one culture to decode data encoded by a single cultural group. Examples include Abelin and Allwood (2000), Beier and Zautra (1972), Van Bezooijen et al. (1983), Van Bezooijen (1984), Dang et al. (2009), Erickson (2006) Graham et al. (2001); Pell and Skorup (2008); Pell et al. (2009a) and Scherer et al. (2001). Since there is only encoded data from a single cultural group available for acoustic analysis in these studies, it is not possible to investigate cross-cultural correlation of vocal cues used.

Some previous cross-cultural studies, such as Anolli et al. (2008) and Fónagy and Magdics (1963) have not included a decoding test, which has meant that the data is not tested to check whether the intended emotion encoded is recognised as that emotion even by speakers from the same culture as the encoders. In other words, there is no confirmation that the vocal features used by the encoders do actually transmit the intended emotion to listeners. Whilst including a decoding test gives a fuller picture, studies which do not include a decoding test are however useful in that they add to evidence of vocal cues which appear to be used in the expression of different emotions and cross-cultural comparison can be made. For example, whilst not including a decoding test which would highlight which items were reliably encoded examples of specific emotions, Anolli et al. (2008), analysed encoded data from more than one culture (Chinese and Italian), making a cross-cultural comparison of acoustic parameters used in the encoded data for each culture and each emotion.

Some previous cross-cultural studies, such as Anolli et al. (2008) and Fónagy and Magdics (1963) have not included a decoding test, which has meant that the data is not tested to check whether the intended emotion encoded is recognised as that emotion even by speakers from the same culture as the encoders. In other words, there is no confirmation that the vocal features used by the encoders do actually transmit the intended emotion to listeners. Whilst including a decoding test gives a fuller picture, studies which do not include a decoding test are, however, useful in that they add to evidence of vocal cues which appear to be used in the expression of different emotions and cross-cultural comparison can be made. For example, whilst not including a decoding test which would highlight which items were reliably encoded examples of specific emotions, Anolli et al. (2008), analysed encoded data from more than one culture (Chinese and Italian), making a cross-cultural comparison of acoustic parameters used in the encoded data for each culture and each emotion.

More recently, Pell et al. (2009b) included emotion vocalisations both encoded and decoded by English, German, Hindi and Arabic participants. Emotions were encoded on pseudo-utterances resembling the participant's native language. As is generally the case, participants who produced the vocalisations were different from those who performed the decoding tests. Burkhardt et al. (2006) investigated emotion vocalisations encoded by computer manipulation of utterances in French, German, Greek and Turkish participants. In these studies, whilst there was no cross-cultural decoding experiment since listeners only decoded vocalisations produced by speakers from their own culture,

the same testing method was applied for each culture which allowed direct comparison to be made in each study between decoding results and vocal profiles for the four cultures investigated.

Galatà and Romito (2010), Pell et al. (2009a), Pfitzinger et al. (2011), Scherer et al. (2001) and Thompson & Balkwill (2006) amongst others have called for 'balanced' or 'symmetrical' studies in this area. This is also evidenced by the quotation at the head of this chapter (Scherer, 2001). A study which incorporates a balanced method would include encoding tests performed by native speakers from different cultural groups and decoding tests which are also performed by native speakers from different cultural groups, and would include both in-group and cross-cultural decoding. This is in line with Scherer's adaptation of The Brunswikian Lens Model (Scherer, 2003), which emphasises the need to consider both the production and perception of vocal cues. This was discussed in Chapters One and Two. However, it has been recognised that the lack of balanced cross-cultural studies has been largely due to the methodological difficulties involved. The construction of a methodology to investigate cross-cultural prosodic cues of emotion which has cross-culturally consistent encoding and decoding procedures presents an even greater challenge, which the present study has aimed to address.

Very few studies have a balanced design in which each culture investigated both encodes and decodes the data, both in-group and cross-culturally. Table 3.2 below shows previous studies in this area which have attempted some element of balance in their design, in which each culture investigated both encodes and decodes the data, both in-group and cross-culturally. Any attempt at symmetry, where a balanced study has a consistent design cross-culturally for both the encoding test and the decoding test, is also shown.

Only two of these studies, Pell et al (2009b) and Abelin and Allwood (2000) include acoustic analysis of the data. However, the study by Pell et al. (2009b) did not include cross-cultural decoding experiments and cannot therefore be considered a balanced study as such and acoustic analysis was performed only on the Swedish data in the experiments by Abelin and Allwood (2000) and Abelin (2004).

Study	Balanced	Symmetrical	Acoustic analysis
Abelin and Allwood (2000) Abelin (2004)	Yes, balanced if include both studies.	No	One language (Swedish) only
McCluskey et al. (1975) McCluskey & Albas (1981)	Yes	Yes, although low-pass filtering could have excluded significant acoustic features.	No
Pell et al. (2009b)	No In-group decoding only	Yes, although 'foreign-sounding' pseudo-utterances could have influenced decoding. Phonetic differences between utterances may have influenced recognition.	Yes
Pfizinger et al. (2011)	Yes	No	No
Sauter & Scott (2007) Sauter et al. (2010)	Yes, balanced if include both studies. However, this is a study of affect bursts, not prosodic cues.	Yes, although this is a study of affect bursts, not prosodic cues	No

Table 3.2: Previous studies which have included some element of balance and/or symmetry in their encoding/decoding method

As can be seen, there are only four previous examples of balanced studies in the area of cross-cultural vocal communication of emotion and one of these studies Sauter and Scott (2007) and Sauter et al. (2010) taken together, investigated affect bursts rather than prosodic cues. This study investigated cross-cultural and in-group recognition of emotion encoded on "raw affect bursts" (section 1.3) such as screams and laughs, rather than vocal signals in speech or pseudo-utterances. Both affect burst and prosodic cues were therefore potentially available to decoders. This study investigated cross-cultural recognition of emotion vocalisations and a search for evidence of universal recognition and did not include acoustic analysis. Pfitzinger et al. (2011) used different testing procedures, collecting Hebrew data from psychotherapy sessions and German data from online gaming sessions to compare native Hebrew and native German recognition of activation, valence and dominance dimensions (see section 3.6 for explanation of these terms). Neither this study, nor the studies by Abelin and Allwood (2000) and Abelin (2004) taken together used the same testing procedure for both cultures investigated.

Of the three balanced studies of cross-cultural comparison of prosodic communication

of emotion, only McCluskey et al. (1975) and McCluskey & Albas (1981) used a symmetrical design in terms of the same testing procedure for both cultures. However, low-pass filtering of the data, in order to mask verbal meaning, could have excluded significant acoustic features in this study.

3.6 Classification of emotion

This section explores how emotion has been differentially classified in previous studies of emotion vocalisation, referring to dimensions, appraisal ratings and discrete categories, the latter being the most common method used, particularly in cross-cultural studies. The discrete categories investigated in cross-cultural studies have tended to be basic emotions as is the case in this study. See Chapter One (Section 1.4) for an explanation of what is understood by the term "basic emotion".

Wundt (1897) is an early example of a work describing emotion in terms of dimensions. Each of these dimensions varies along a continuum. Some more recent studies refer to dimensional characteristics of emotion, although the labels attached to these dimensions vary between researchers. The dimension most commonly referred to is arousal (also termed activation or activity), which refers to level of arousal or relaxation represented by a particular emotion. Another dimension sometimes referred to is valence (also termed evaluation), which refers to how pleasant/positive or unpleasant/negative an emotion is. A further dimension, which is less commonly referred to, is control (also termed power or potency or dominance), which indicates whether an emotion represents greater or less dominance. For example, Angry would be classified as aroused, unpleasant and dominant.

The dimensional approach has been used particularly in studies of vocal emotion applied to speech synthesis and human-machine interaction, such as Cowie et al. (2001) who investigate emotion in terms of evaluation, activation and power. A recent cross-cultural study of vocal emotion (Pfitzinger et al., 2011) also takes a dimensional approach, referring to activation, valence, and dominance. For an overview of the literature leading to consensus on conceptualization of emotion along three dimensions, see Schröder (2004).

Scherer's component process model (Ellsworth & Scherer, 2003; Scherer, 1984; 1986; Scherer & Ellgring, 2007) relates specifically to the relevance of appraisal theory to vocal emotion, and predicts vocal changes relating to particular emotions. Most of these predictions were verified by Banse and Scherer (1996). According to the component process model an individual makes a subjective appraisal of the significance for that individual of a stimulus by performing recurring stimulus evaluation/appraisal checks (SECs), which result in the experience of an emotion. It is suggested that individuals use certain evaluative criteria to subjectively evaluate or appraise events, checking the stimulus they are experiencing thus eliciting an emotion. These evaluative criteria are also referred to as dimensions, not to be confused with the different use of this same term above. Evaluative criteria or dimensions include an individual's appraisal/evaluation in terms of 'novelty' (how suddenly and abruptly the event occurred), 'intrinsic pleasantness' (how pleasant they rated the event), 'goal significance' (whether the person appraised that the event was more or less likely to be conducive to reaching a goal or satisfying a need) and 'urgency' (whether the person appraised that they needed to respond to the event urgently). The individual would also appraise their 'coping potential' in terms of their own and/or others' 'responsibility' for and 'control' over the event and 'norm compatibility', that is whether or not the event was compatible with the individual's norms or standards. The model predicts that the outcome of this appraisal process directly determines physiological response, motor expression and preparation for response. For example, Scherer and Ellgring (2007) predict the following multi-modal example:

> ...anger is expected to be the result of an event being appraised as an obstruction to reaching a goal or satisfying a need, produced by an unfair intentional act of another person, that could be removed by powerful action (with a correspondent response patterning consisting of aggressive action tendencies, involving sympathetic arousal, knitted brows, square mouth with teeth clenched, and loud, strident vocal utterances) (p.159).

Laukka et al. (2010) is one of the few cross-cultural studies to have incorporated this approach. This study analysed the ability of speakers of US English to rate emotions expressed by speakers representing five different varieties of English, including US English.

Jaywant and Pell (2012) found that "despite sharing the same valence, anger, sadness, and disgust are communicated and recognised as unique emotion categories" (p.9).

Cross-cultural studies of the expression of human emotion have tended to conceptualise emotion in terms of discrete emotions, particularly basic emotions, as is the case in the present study. This follows on from facial expression research (Ekman & Friesen, 1971; Izard, 1971) which, as discussed in Chapter One, has suggested evidence of universals in facial expressions relating to what are regarded as basic emotions (Happy, Sad, Angry, Fear, Surprise and Disgust). Most cross-cultural studies of vocal emotion, including the present study, have conceptualised emotion in terms of these discrete categories. Four of these categories (Happy, Sad, Angry, and Fearful) in particular tend to be investigated since these are the categories for which most evidence of universals has been found in facial expression research. Whilst these basic emotions are not prevalent in natural speech (Douglas-Cowie et al. 2003), they have served as a useful basis for most cross-cultural research of vocal emotion, including the present study.

Where studies of the vocal communication of emotion have conceptualised emotion in terms of discrete categories, emotions included have varied from one work to another, although what psychologists have labelled as 'basic' emotions (see 1.4) are those most often studied. Investigation of the existence of possible universals in vocal expression of emotion has tended to focus on basic emotions as these are the emotions which, if any, would be most likely to exhibit universal tendencies, given evidence of phylogenetic continuity and facial expression research (see Chapter One). Studies most commonly include investigation of two or more of Happy, Sad, Angry, Fear, Surprise and Disgust, the first four being the most commonly studied and are therefore those for which the most evidence of cross-cultural similarity has been found in their vocal expression. The cross-cultural study by Anolli et al (2008) and that by Van Bezooijen (1984) also included investigation of Contempt.

Cross-cultural studies have occasionally included emotions which have been referred to as 'secondary' or 'social' emotions: Anolli et al. (2008) included pride, guilt and shame and Nakamichi et al. (2002) included sarcasm in their study. Gobl and Chasaide (2003) include relaxed, stressed and bored which they categorise as moods and formal, interested and friendly which they class as speaker attitude.

Previous studies of emotion vocalisation have sometimes included a 'Neutral' category as a kind of reference point or baseline against which to measure variation due to different emotions, it being assumed that emotion involves some kind of 'visceral perturbation'.

3.6.1 *Cultural bias of emotions studied and emotion labels*

In cross-language studies, a complicating factor is the translation of emotion labels. Monolingual cross-cultural groups and even different individuals or the same individual at different points in time, may attach different connotations to the same emotion label. It is also possible that there is a Western cultural bias as to which emotions are regarded as basic and further research in this area may well be helpful.

Therefore, another complication for researchers setting up cross-cultural research in this area is the controversy which exists regarding the possible cultural bias attached to the emotion words used in these studies. Murray and Arnott (1993) commented that definitions of different emotions may have tended to be biased towards Western cultures. See also Wierzbicka (2009) for further discussion. It is difficult to know whether the emotion being expressed is actually the same emotion in both cultures. Speakers from different cultures and even different individuals within the same culture or the same individual at different points in time, may also attach different connotations to the same emotion label. If there are any quasi-universally common emotion categories, these are most likely to be basic emotions, as found in research on facial expression, since these emotions may have phylogenetic continuity in terms of psycho-physiological influence.

3.7 Methods used to encode vocal emotion

There is also controversy as to which method is most appropriate for encoding data. Possible methods include natural, induction, computer synthesis and computer manipulation, simulation by human portrayal and multimodal studies.

Most previous research, particularly cross-cultural studies, have tended to use simulated data, where emotion portrayals are made by participants, sometimes professional actors, in an encoding experiment. 87% of all studies of vocal expression of emotion which were included in the meta-analysis by Juslin and Laukka (2003) used this method and in 13% of these masking was used (see 3.8). Cross-cultural studies continue the tendency to use data simulated in emotion portrayals.

The following sections explain the trade-off between consistency and authenticity and the advantages and pitfalls of the various methods of data collection used in previous studies of the vocal expression of emotion.

3.7.1 *Ecologically valid data*

Natural, spontaneous speech encodes the influence of social convention upon emotion vocalisation. However, whilst this data may be regarded as ecologically valid, just because a speaker produces vocal cues suggestive of a particular emotion does not necessarily mean that the vocalisation is authentic in the sense that the speaker experienced the emotion recognised by listeners. The speaker may not actually be experiencing the emotion they are vocalising, and as Scherer (2013b) argues, the determination of what is the "true" underlying emotion of the encoder is a challenge unlikely to be met by the current research options available and is a challenging area for future research.

Finding examples of 'natural' vocalisations of emotions, especially basic emotions, can also be difficult. Only 12% of studies included in Juslin and Laukka's meta-analysis used natural speech samples and most of these were studies of Fearful expressions which were made during aviation accidents. One early study (Williams & Stevens, 1972) made an acoustic analysis of the recording of the stressed speech of a radio announcer reporting on the Hindenberg zeppelin disaster as it happened. This study did not include a decoding experiment.

Further examples of monolingual studies using natural data can be found in reviews by Murray and Arnott (1993); Juslin and Laukka (2003); Douglas-Cowie et al. (2003); Ververidis and Kotropoulos (2003); Ververidis and Kotropoulos (2006); Scherer (2003).

It is even more problematic to access comparable natural, spontaneous examples of vocalisations of emotions in different cultures in order to make a cross-cultural comparison of recognition of emotion from the voice and any acoustic patterns of emotion. Spontaneous speech data has been used in very few cross-cultural studies on

vocal emotion (Chung, 2000; Scherer and Ceschi, 2000; Erickson, 2010; Pfizinger et al., 2011). Scherer and Ceschi (2000) gathered data at the lost baggage claims office at Geneva airport International Airport from passengers who had found that their luggage had failed to arrive. Emotions included were 'Anger', 'Good Humour', 'Indifference', 'Stress' and 'Sadness'. As one might expect, the vocalisations gathered were mainly of Frustration and Anger. Chung (2000) used a method of data collection which allowed the researcher to retain a certain amount of control. Participants were interviewed about times in their past when they may have felt particular emotions. It may be argued that this forms a kind of encouragement to self-induction. Pfitzinger et al. (2011) collected data from psychochtherapy sessions for Hebrew and from online gaming sessions for German.

A balanced, fully symmetrical cross-cultural study would not be possible using natural data unless the data is induced, in which case it may also be more feasible to monitor whether or not the vocalisations expressed 'true' feelings. In addition, some or all emotions may be more or less likely to be expressed in different cultures within different contexts.

3.7.2 *Induction and self-induction*

Only 7% of studies in Juslin and Laukka's meta-analysis (2003) reported using an induction technique (e.g. Skinner, 1935; Sobin & Alpert, 1999). All of these were mono-cultural studies, mostly of English. A more recent mono-cultural study which compares induction and portrayal techniques is that by Scherer et al. (2011).

Techniques used to induce emotion in empirical studies in psychology include in particular the Velten induction procedure (Velten, 1968; Kenealy, 1988; Wilting et al., 2006), in which a participant is asked to put themselves into the particular emotional state as far as possible and to read aloud statements which verbally express the specific emotion. This technique is related to the Method acting technique, largely influenced by the Stanislavski 'system' developed in the 1920s (Stanislavski, 1967), in which actors use memory to 'relive' their own physical and psychological experiences of emotion in order to 'experience the part' and improve their portrayals. A tendency towards hysteria in some actors using this technique led Stanislavski to advocate the use of the imagination and a belief in the emotion within the circumstances of the text to aid

emotion portrayals, rather than memory. This revised method was adopted by many proponents of Method acting, including Adler (1988) but not by others, such as Strasberg (1987) who continued to advocate the use of memories of emotional experience to improve portrayals.

Although it has been claimed that induced emotion data will be more authentic than simulated data, it is recognised that there can be ethical constraints, particularly in the induction of negative emotions. The same stimulus may also induce different emotions in different cultures or in different individuals and the induction of emotion or particular methods of induction may be more or less appropriate or tolerated in some cultures than in others. It could be argued that sometimes there is not such a clear dividing line between the induction method and the simulation method and the method of collecting spontaneous data from a real setting. Acted emotions may result in the self-induction of emotion and where a participant experiences the emotion, this may be regarded as authentic data, even though the data was not gathered in a real setting.

It is possible that acted portrayals in simulation procedures employed in studies of the vocal correlates of emotion (see 3.7.3) may result in the experience of the acted emotions, depending upon the individual and any acting technique used; for example where Stanislavski or Method acting techniques are used by the participant. In other words, acted vocal portrayals of emotion may lead to what could be argued to be 'self-induction' of emotion which would avoid possible ethical problems which may be associated with external induction. For example, in the cross-cultural study by Anolli et al. (2008), participants encoded emotion by reading stories which were intended to express particular emotions, reading these as convincingly as they could. Whether or not the participant actually felt the emotions they vocalised is not recorded.

Data which has been obtained through induction may also be classed as real data in the sense that the expressions are produced with authentic emotion. For example the Reading–Leeds database, reported in Roach et al. (1998) and Douglas-Cowie et al. (2003), was obtained by conducting radio and television interviews in which speakers were induced by interviewers to relive emotionally intense experiences. Since this data represents authentic emotion, it may also be argued that it is real or natural.

There is evidence that music can induce emotion (Scherer et al. 2002; Scherer, 2004) and that music in a major key can express and induce positive valence emotion

(Webster & Weir, 2005; van der Zwaag et al., 2009), whilst slow music tends to express and induce low arousal emotion (Pinchot-Kastner & Crowder, 1990; Webster & Weir, 2005). It has also been suggested that there is sensitivity to emotion in music from a culture which is not one's own (Balkwill & Thompson, 1999). However, empirical studies to date using musical induction techniques tend to be within the area of clinical psychology. (Clark, 1983; Juslin and Sloboda, eds., 2001; Kenealy, 1988). Scherer (2013b) compared the acoustic features of induced emotion data (Happy and Sad) with data obtained by actor portrayal. Actor portrayed data is sometimes criticised as being artificially exaggerated. However, Scherer (2013b) found that the acoustic characteristics of induced and acted data showed strong similarities and that if anything, the more 'authentic' induced data exhibited more pronounced differences between acoustic patterns for the two emotions studied.

It is argued that the distinction drawn between spontaneous data gathered from a normal setting, simulated emotion and induced emotion is not necessarily clear-cut. This blurred distinction was utilised in the procedures constructed to elicit vocal expressions of emotion in the present study.

3.7.3 *Simulation by human portrayal*

Cross-cultural research has tended to use simulated data rather than natural data partly in order to facilitate greater cross-cultural consistency in the stimulus and context across the cultures being investigated in any particular study. In addition, one main aim of cross-cultural research has been to investigate possible universal tendencies in emotion vocalisation, following on from research which found universals in the facial expression of emotion (Ekman & Friesen 1971; Izard, 1971; Matsumoto, 1996) as discussed in Chapter One. The frequency with which basic emotions occur and the context in which they occur may vary cross-culturally. Simulated data has also been used in order facilitate greater cross-cultural consistency in the stimulus, context and materials used.

In studies which include vocalisations by participants from different cultures, simulation of basic emotion vocalisations has allowed the researcher greater control to design experiments which use the same stimulus for production of emotion for all participants, including reading passages and translations of these and emotion labels in the language of the speaker, although there is still the possibility that participants from different

cultures and even participants from the same culture may interpret the same labels differently. Simulation has served as a useful basis for most cross-cultural research of vocal emotion.

Simulated data has therefore often been used to give greater consistency and in an attempt to avoid the ethical problems of the induction method. Nevertheless, it is possible that simulating emotion may actually lead to the experience of the emotion by feedback response.

Cross-cultural studies have tended to use this method in order to increase cross-cultural consistency in the verbal channel, experimental context and the emotions expressed, although, given problems of translation of emotion labels and utterances (3.6.2), potentially different conceptualisation of emotions and culture-dependent interpretation of experimental context, this consistency can certainly not be guaranteed.

Where emotions are simulated, various methods have been used. Speakers have sometimes been professional actors, sometimes not. In the study by Anolli et al. (2008), forty-eight undergraduates (29 Chinese and 19 Italian) were asked to read aloud short stories inducing different emotions (joy, sadness, anger, fear, contempt, pride, guilt, and shame) within a scenario approach in which emotionally-charged verbal utterances were embedded in reading passages. No decoding test was conducted so no masking of the verbal channel was required. Acoustic analysis was carried out on the utterances.

Utterances which the researcher regards as semantically neutral may be presented to the speaker who is asked to produce the utterances with different emotions. The rationale for using specially constructed semantically neutral utterances has been that it becomes unnecessary to mask the verbal channel where a decoding test is performed. Where verbal utterances are not semantically neutral and a recognition test is performed, there is a need to mask the verbal channel and masking methods are discussed below. Where the decoding test is performed by participants who do not speak the encoder's language, masking has not been necessary. However, where a study aims to test native and non-native decoding accuracy, masking of the verbal channel has been conducted in order not to give native speakers an advantage from verbal cues to the emotion being expressed.

Mono-cultural studies have often used simulated data (Murray & Arnott, 1993). For further examples, see Juslin and Laukka (2003). A large-scale cross-cultural study using

simulation to encode emotion was reported in Scherer, Banse and Wallbott (2001). As mentioned above, native German speakers simulated emotion on pseudo-utterances and recognition accuracy was tested for speakers from nine cultures.

It has been suggested that simulated data may not be representative of how everyday emotions are expressed in speech. Eliciting authentic, natural data in studies of emotion vocalisation is vital if we are to understand how the voice communicates emotion in everyday speech. Studies focussing specifically on cultural influence upon emotion vocalisation and the expression of emotion in everyday speech require data which is as natural as possible which will incorporate display rules, including masking. In addition, as Greasley et al. (2000) points out, it is often the case that more than one emotion may be encoded in a short stretch of natural speech. Scherer (1986) comments that simulated emotion vocalisations are dependent upon acting abilities of participants and possibly on the acting techniques they use, commenting that "unless actors are given detailed scenarios, they may use very different approaches to producing the appropriate expression...thinking of a personal experience...imitating cultural stereotypes, among others." (Scherer, 1986 p.146)

Some studies have tested the comparative recognition of acted and spontaneous speech. Scherer (2003) argues that the main difference between acted and spontaneous speech is that acted expressions are more aroused, exaggerated or stereotypical. Laukka et al. (2007) and Wilting et al. (2006) found that the intensity of emotion was rated higher by listeners where utterances were acted compared to spontaneous utterances. Audibert, Aubergé and Rilliard (2008, 2010) analysed corpora of acted and spontaneous expressive speech and found that listeners could often distinguish between acted and spontaneous speech but that emotion discrimination was generally not significantly affected by whether the speech was acted or spontaneous and no significant effect of the emotional category was found when comparing acted and spontaneous speech. Further research is required comparing recognition rates and vocal cues of acted and spontaneous speech.

It may be argued that there is not necessarily such a clear dividing line between natural and simulated data where portrayal may result in self-induction. Scherer (2013b) also makes this case. This is not to suggest that simulated data which may sometimes cause the speaker to self-induce emotion may be regarded as natural data in the same way as that found in spontaneous everyday speech but that self-induction may have a part to

play in providing access to data, particularly for basic emotions which are seldom found in everyday speech. This method of data collection can also be useful for symmetrical cross-cultural studies where more control may be required in the semantic and phonotactic characteristics of utterances from more than one culture.

Whilst most previous cross-cultural studies of the vocal expression of emotion, use acted data, it is possible that at least some of the participants may have self-induced the emotions they vocalised at least some of the time perhaps due to a feedback effect. However, it is not possible to ascertain which of the data is artificial and which is authentic. This is also the case in the present study.

Scherer (2013b) found that vocalisations of induced emotions (Happy and Sad) exhibited more exaggerated acoustic patterns than that of acted data. As discussed above, emotions were induced by using the Velten technique with accompanying music corresponding to the emotion. For further discussion of the relative merits of using natural, induced and simulated data in studies of the vocal communication of emotion, see Cowie et al. (2005) and Scherer (2013b).

3.7.4 *Synthesis and computer manipulation*

One of the ways in which people have conducted experiments has been to synthesise data to test for the significance of specific acoustic variables in the vocal communication of emotion. A recent study by Yanushevskaya, Chasaide, and Gobl (2011) investigated the role of voice quality and specific features of f0 which signal emotion. Utterances were synthesised with different voice qualities and Irish-English, Russian, Spanish and Japanese listeners were asked to rate the utterances for affect. The authors use the term affect to include not only emotions (Happy, Sad, Fearful), but also attitudes (apologetic-indignant, bored-interested, intimate-formal, relaxed-stressed, and fearless).

Burkhardt et al. (2006) systematically manipulated semantically neutral utterances in terms of f0, duration and jitter to test the relevance of each of these parameters in the recognition of emotions.

The data in Pell and Skorup (2008) was encoded on pseudo-utterances the duration of which was then manipulated and a test was created to ascertain whether the duration of

Happy and Sad utterances significantly affected recognition accuracy.

Speech data collected from ecologically valid settings has also been used as input to create concatenated speech, which synthesises emotional speech to aid human-computer interaction. In concatenative synthesis, the resulting synthesised speech will lose its naturalness where phonetically and prosodically unmatched speech sections are strung together. However, research in this area is striving to meet the substantial challenge of improving naturalness in synthesised speech. For reviews of issues and approaches in this field, see Cowie (2009); Douglas-Cowie et al. (2003); Murray and Arnott (1993), Petta, Cowie and Pelachaud (2011) and Schröder (2001). The Belfast Naturalistic Database (Douglas-Cowie et al. 2000; Douglas-Cowie et al., 2003) consists of data taken from ecologically valid television current affairs programmes and chat shows as well as interviews conducted by the research team. The data in the Belfast Naturalistic Database was rated by three trained raters, according to two dimensions, activation and evaluation, using a computer programme, known as Feeltrace. More information about this system can be found in Cowie and Cornelius (2003).

3.7.5 *Multimodal studies*

The communication of emotion tends to be a complex multichannel process, incorporating vocal, verbal, facial and body expression and gestures (Scherer & Ellgring, 2007). Studies which look at more than one channel through which cues can be transmitted are known as multimodal studies. Multimodal research has been gaining increasing focus, particularly in the search for appropriate databases for application in human-computer interaction (e.g. Douglas-Cowie et al., 2005).

Studies by Pell and colleagues (Pell, 2005a; 2005b; Pell et al., 2005; Pell & Skorup, 2008; Paulmann & Pell, 2010 and Rigoulot & Pell, 2012) used different variants of what they refer to as a Facial Affect Decision Task (FADT). For example, in Pell and Skorup (2008), the duration of pseudo-utterances produced by native speakers of Arabic were synthetically manipulated to lengths of 66ms and 1000ms to test for the influence of pseudo-utterance duration on the recognition of congruous and non-congruous Happy, Sad and non-emotional static facial expressions. Decoder participants were 50 native speakers of Canadian English.

Rigoulet and Pell (2012) investigated the influence duration of emotional prosody duration upon the length of time and frequency with which congruent or incongruent facial expressions were looked at and found “an emotion congruency effect” since where facial expressions were congruent with the emotional prosody, faces were looked at for longer regardless of length of utterance.

The mono-cultural study by Jaywant and Pell (2012) also used the FADT. In this study 50 English listeners rated pseudo-utterances and static emotional face targets as Angry, Sad, Disgusted or Neutral. The results of FADT studies demonstrate that facial expressions of emotion are judged more quickly and more accurately when they are preceded by emotional prosody which is congruent rather than incongruent with the facial expression.

Little work has been done so far on the trade-off between different modalities in the communication of emotion and the focus in the present study is on vocal cues. However, the interplay of different modalities is recognised in this study, for example in the consideration of the facial feedback response (1.5.1).

3.8 Isolating vocal cues from verbal cues

An important methodological challenge in the design of experiments to investigate the vocal communication of emotion is how to isolate vocal cues from verbal cues since where verbal cues are present, these may influence the emotion recognised by the listener, whether the emotion communicated by verbal cues is the same as, or conflicts with, vocal cues.

Electronic masking procedures have also been used to attempt to mask the verbal channel. Ishii et al. (2003), for example, used simulated data, asking speakers to produce any two sentences they wished in their native languages. The researchers then filtered out the verbal content electronically by low-pass filtering. McCluskey et al. (1975), Albas et al. (1976) and McCluskey and Albas (1981) also used low pass filtering to isolate the vocal channel from verbal cues. However, all electronic masking procedures can pose further problems. A low-pass filter removes higher frequencies which may be important for the expression of emotion. Speech may be played backwards; however, this distorts intonation contours. Random splicing removes pauses

and reorders the remaining recording but this removes temporal features and distorts intonation. The use of a throat microphone and laryngograph is not completely effective in masking the verbal channel and does not pick up supra-laryngeal activity. Semantically neutral utterances and synthesised data have also been used to mask verbal content as discussed in sections 3.7.3 and 3.7.4 respectively. Some studies have encoded emotion on affect bursts (Schröder, 2003; Sauter & Scott, 2007; Sauter et al., 2010). See 1.3 for an explanation of affect bursts.

Ohala (1984) argued that there may be phoneme-specific segmental influence, particularly from vowels upon the recognition of particular emotions. The present study takes some account of this possibility in relation to vowels in the design of the encoding experiment.

Pseudo-utterances, sometimes referred to as nonsense utterances, have also been used to isolate the vocal channel (Rigoulot & Pell, 2012; Pell & Kotz, 2011; Pell et al., 2009a; Pell et al., 2009b; Scherer et al., 2001). These pseudo-utterances aim to exclude all verbal meaning and tend to be constructed to mimic the phonotactic properties of the encoder's language and sometimes the decoder's language(s). The use of pseudo-utterances allows emotion judgements to be based on vocal parameters only.

An alternative method used is for emotions to be vocalised in a language which participants in a recognition experiment do not understand. However, it is possible for example, that certain sounds present in the verbal channel of a language may carry affect significance for speakers of another language which does not contain these sounds.

However, in symmetrical studies where listeners from each culture are asked to perform decoding tests on the data, whilst all listeners will have access to vocal cues to emotion, those who understand the verbal content will also have access to possible verbal cues. Pfitzinger et al. (2011) for example, included verbal data from both Hebrew and German native speakers and decoding tests were performed by both Hebrew and German listeners. This meant that possible verbal cues were present for one group or the other for each item. Pfitzinger et al. (2011) suggest for example, that "the reason for the lower correlation between German and Hebrew judgments of Hebrew emotional speech" (p. 1589) could be "the higher verbal content... of the emotional content in the Hebrew data" (p. 1589).

3.9 Findings of previous studies

This section discusses the findings of previous studies and summarises questions which are raised and which are relevant to the present study and informed the hypotheses and predictions made.

3.9.1 *Recognition accuracy*

Since studies vary according to encoding and decoding methods used, the number and profile of encoding and decoding participants and the number and combination of emotions investigated, comparison between specific levels of decoding accuracy between one study and another needs to be treated with caution. However, previous studies have tended to find that decoding accuracy occurs at levels well above chance for both native and non-native decoders. Whilst most studies have investigated cultures represented by Indo-European, particularly Western European languages, this has recently begun to change with research also starting to focus upon cultures represented by other languages.

This provides evidence of pan-cultural psycho-physiological influence upon the vocal communication of emotion and supports Scherer's (1986) suggestion that acoustic cues of emotions exist which are psycho-physiologically influenced (covariational effects) as discussed in Chapter One. Scherer, Banse and Wallbott (2001) for example, collected vocal expressions of Joy, Sad, Anger, Fear and Neutral on pseudo-utterances by native German speaking actors. They tested recognition of these expressions by listeners from nine countries (Western and non-Western) in a forced judgement test and found an overall level of accuracy of 66% across all countries and all emotions although the level of decoding accuracy varies between one emotion and another as discussed in section 3.9.4. Thompson and Balkwill (2006); Anolli et al. (2008), Pell and Skorup (2008) and Pell et al. (2009a; 2009b), for example, have also all found recognition levels far greater than would have been expected by chance.

In their review of 97 studies of visual and prosodic emotion cues, Elfenbein and Ambady (2002, p.204) stated that "…although emotions are recognised at above chance levels across cultures, there is also cross-cultural variation in recognition accuracy"

Sauter et al. (2010) found that non-verbal vocal signals such as screams and laughs, representing basic emotions (Joy/Amusement, Sad, Angry, Fearful, Disgusted and Surprised) were reliably identified cross-culturally across two highly unrelated cultures (British English and Himba). The authors argued that this provided evidence to suggest "that vocal signals of emotion are, like facial expressions, biologically driven communicative displays that may be shared with nonhuman primates." (Sauter et al., 2010, p.2410).

Variation in recognition rates between one culture and another may be partly explained by differences in display rules between one culture and another. The concept of display rules was discussed in Chapter One. Sadfar et al. (2009) compared 'emotional display rules' for seven basic emotions within and across three cultures: Japanese, US American and Canadian. The results of this study suggested that Japanese displayed less expression of high power emotions (Anger, Disgust, Contempt) than did the other two groups. Positive emotions (Happy, Surprise) were also significantly less expressed by Japanese participants than by Canadian participants. It is possible that if a culture displays significantly less expression of particular emotions or of emotion in general, this may influence the recognition of particular emotions or emotion in general when emotions are expressed by other cultures. Finding evidence as to what the vocal cues are and how they may be combined to form vocal profiles, which discriminate between distinct emotions, clearly requires acoustic analysis of the encoded data. Previous research in this area is discussed in Chapter Two.

3.9.2 *In-Group Advantage*

Evidence of In-Group Advantage has been found in previous studies: participants of the same culture as the encoder tend to be more accurate at recognising emotion vocalisations than are participants from a different culture. However conflicting results have been found in relation to the Cultural Proximity Hypothesis (3.9.3), according to which the closer a culture is to one's own, the more likely one is to recognise emotion vocalisations by participants from the related culture.

Frick (1985) noted that of the very few cross-cultural studies made to date, some found that cross-cultural recognition was as accurate as monolingual recognition. Where others found that cross-cultural recognition was adversely affected, this phenomenon

was termed ‘ethnic bias’ (Kilbride & Yarczower, 1983), later referred to as ‘In-Group Advantage’ by Elfenbein and Ambady (2002). The concept of In-Group Advantage was later extended to include the possibility that cultures which are more similar in terms of, for example, type of power structure, degree of individualism or collectivism, are more likely to identify each other’s vocally expressed emotion than listeners from more different cultures. This has been referred to as the ‘Cultural Proximity Hypothesis’. (Elfenbein & Ambady, 2003). Findings related to the Cultural Proximity Hypothesis in previous studies are discussed further in the following section (3.9.3).

Thompson and Balkwill (2006) found that whilst English decoders identified Joy, Sad, Anger and Fear at an accuracy level above chance in items encoded by native speakers of English, German, Chinese, Japanese and Tagalog, evidence of In-Group Advantage was also demonstrated since English listeners’ overall level of accuracy was greater for recognition of the English data.

However, whilst evidence in support of In-Group Advantage has tended to be reported in previous studies, this is not always the case. For example, Erickson (2010) found that Japanese and American English listeners identified emotion vocalisations by a Korean speaker more accurately than Korean listeners identified the emotions in the same vocalisations. The author suggested that native listeners may attempt to include linguistic processing of the vocalisations whilst non-native listeners would expect to have to make judgements on non-linguistic information only. The author indicates that most Korean listeners reported that their performance may have been affected by the fact that they knew that they were listening to Korean so may have been expecting lexical cues to the emotions which were not generally evident on the short vowels and single words with which they were presented. A possible solution suggested by the Erickson (2010) would be to present the data to Korean listeners without informing them that the vocalisations were Korean.

Erickson (2006) also appears to contradict the concept of In-Group Advantage since American English listeners identified Japanese expressions of Happy, Sad, Angry, Surprised and Suspicious vocalised on the Japanese word “banana” more accurately than native Japanese listeners recognised the same vocalisations. Korean listeners also had higher recognition rates than native Japanese listeners for each emotion except for Japanese Happy and identified Japanese Angry at a higher level of accuracy (100%) than either Japanese or American English listeners. The authors suggest that whilst

Japanese listeners may have been distracted from prosodic cues by "an anomaly between lexical meaning and prosodic meaning" (Erickson, 2006, p.4), this may not have been the case for American English listeners as they may not have recognised the Japanese pronunciation of the loanword "banana" since in Japanese the second syllable is not stressed and the initial and final vowels are not weakened.

Sauter et al. (2010) found that native Himba speakers were less accurate overall in their recognition of emotion from vocal signals such as laughs and screams than were European native English speakers whether in-group or cross-cultural decoding was observed. However, the authors suggested that this was probably due to the English group being more familiar with psychological testing than the Himba group.

More recently, a few studies have also suggested that some cultures may decode vocally expressed emotion less accurately than others. Several explanations have been put forward for this. For example, Kitayama and Ishii (2002) suggested that cultures may differ in their reliance on vocal or verbal cues to identify emotion. They found that Japanese decoders may rely more on vocal cues than US English listeners who, they suggested, relied more on verbal cues. Erikson (2006) found that whilst Japanese, Korean and US English listeners recognised vocal expressions of emotion by native Japanese speakers at levels well above chance, US English listeners were the most accurate, followed by Korean. Japanese listeners were the least accurate perceivers of vocalisations by Japanese speakers. The authors suggested that this may be due to the Japanese participants recognising an anomaly between the prosodic cues and lexical meaning of the one-word utterance (the Japanese word "banana"), whilst the US English listeners may not have recognised the loanword due to the different stress pattern used in Japanese.

The possibility that some cultures may rely more than others on vocal cues for the expression and perception of emotion may add further complexity to consideration of In-Group Advantage and the Cultural Proximity Hypothesis. More recently, some evidence has been found some cultures have decoded vocally expressed emotion less accurately than others even when decoding vocalisations by speakers from the same cultural group as themselves (Erikson, 2006; Sauter et al., 2010).

Conducting symmetrical studies such as the present study may help to clarify the aspects discussed above by testing whether or not vocal emotion is more accurately

identified by in-group or cross-cultural decoders, the extent to which this may be the case and whether or not In-Group Advantage is emotion dependent. Symmetrical studies may also help in gaining a clearer picture of possible cross-cultural differences in reliance on vocal compared to verbal cues for the expression and perception of emotion.

3.9.3 *The Cultural Proximity Hypothesis*

Related to In-Group Advantage, it has also been suggested (Elfenbein & Ambady, 2003) that there is cultural proximity advantage wherein the closer the decoding culture is to the encoding culture, the higher the recognition accuracy. Scherer et al. (2001), included data encoded on pseudo-utterances by German actors and decoded by native speakers of languages from seven countries in Europe (German, Dutch, British English, Swiss French, Italian, Spanish) one from Asia (Bahasa Indonesian) and the US (American English). They found that whilst decoding accuracy was well above chance for all cultures (section 3.9.1), the recognition rate was most significantly poorer for Indonesian listeners, thus supporting the Cultural Proximity Hypothesis, if we associate dissimilarity in language typology with cultural difference as Bahasa Indonesian is the only language in the group which is not an Indo-European language.

Sauter et al. (2010) found that basic emotions (Joy/Amusement, Sad, Angry, Fearful, Disgusted and Surprised) were reliably identified cross-culturally across two highly unrelated cultures (British English and Himba). Sauter and Scott (2007) found that positive emotions (achievement/triumph, relief, sensual pleasure), expressed in sounds such as amused laughter and relieved sighs, were discriminated well within group and cross-culturally and found evidence of In-Group Advantage. However, Sauter et al. (2010) found that these positive emotions were not very well distinguished cross-culturally. The discrepancy in results between the two studies could be explained by the Cultural Proximity Hypothesis since Swedish and British English cultures (Sauter and Scott, 2007) are closer to each other and have had far more contact than Himba and British English (Sauter et al., 2010).

Pell et al. (2009a) found that Argentine Spanish decoders identified basic emotions at a level well above chance, whether these emotions were produced by Argentine Spanish, Arabic, English or German speakers, at 64%, 59%, 58% and 56% respectively,

identification accuracy being only slightly improved when recognising Spanish vocalisations of emotion. Listeners performed a forced judgement test which included six alternative responses - anger (enojo), disgust (repugnancia), fear (miedo), sadness (tristeza), joy (alegria), and neutral (neutralidad). According to Pell et al. (2009a), their results did not support the Cultural Proximity Hypothesis since the Argentine Spanish listeners did not identify vocal emotions significantly more accurately from English and German data than the Arabic data even though Arabic is more linguistically distant from Argentine Spanish, all languages being Indo-European apart from Arabic which is Semitic.

Pell et al. (2009a) comment that "Contrary to expectation, our comparisons supply no evidence that language similarity was an overall predictor of how Spanish participants recognised vocal emotions in the cross-cultural setting…" (p.117). The authors therefore concluded that "…linguistic similarity is not a consistent factor which predicts the accuracy of vocal emotion recognition across languages." (p.117), suggesting that greater recognition accuracy scores may be obtained between languages grouped according to "intonational or timing properties… rather than by language typology or families…" (p.117). They comment:

> It is possible that, if language similarity were defined according to specific intonational or timing properties of a language, rather than by language typology or families as is common in the literature, future investigations would be better positioned to evaluate the importance of this variable and its relationship to the in-group processing advantage during vocal emotion processing (p.117).

Thompson and Balkwill (2006) found that English speakers' overall level of accuracy in recognition of emotions in German, a related language, was 67.5% and in Tagalog, an unrelated language, was greater, at 72.2% which appears to contradict the Cultural Proximity Hypothesis if we take language typology as a basis for determining cultural similarity. However, the native speakers of Tagalog in Thompson and Balkwill's (2006) encoding experiment were from the Philippines where English has also been an official language since 1935. The two languages are sometimes so merged in daily life in the Philippines as to have developed the labels Taglish and Englog. Code-switching between Tagalog and English is very common and it is therefore quite possible that English may have influenced Tagalog speech, possibly also influencing the expression of emotion in speech. This may possibly have helped British English listeners in the decoding of emotion from Tagalog speech.

3.9.4 *Recognition rates for specific emotions*

In previous studies, Sad and Angry have generally been found to be most accurately identified from vocal expression. Happy vocalisations are reported to be less accurately identified cross-culturally. This contrasts with facial expression research which found Happy to be the most accurately identifiable. Differing levels of recognition of specific discrete emotions and possible explanations are discussed here.

Whilst Happy faces tend to be the most accurately identified of the basic emotions cross-culturally (Ekman, Sorenson, & Friesen, 1969; Izard, 1994), studies of vocally expressed emotion are tending to find that Sad and Angry are much more accurately recognised cross-culturally than Happy and Fearful and that Happy is often the most difficult basic emotion to identify in the voice cross-culturally (see Juslin & Laukka, 2003). This supports the usefulness of multimodal communication. Scherer et al. (2001, p.89) commented that "One of the reasons may be that happiness or joy is strongly marked by smiling, the ubiquitous activity of the zygomaticus muscle, in joy-related emotions. No comparably iconic cue may exist for the voice". (Scherer et al., 2001, p.89)

As specific examples for the voice, Erickson (2010) found that Japanese, Korean and American English listeners identified Anger and Sadness vocalised by a Korean speaker more accurately than they identified the speaker's Happy vocalisations and Abelin and Allwood (2000) found that Anger and Sadness vocalised by a Swedish speaker were more accurately identified than Joy, Fear, Surprise, Disgust, Shyness and Dominance by Swedish, English, Finnish and Spanish listeners.

In contrast, Pell et al. (2009a) claim there is evidence to suggest that decoders from the same cultural group as encoders recognise vocalisations of Joy more accurately than Sadness, Anger, Fear or Disgust expressed vocally. As discussed in 3.9.2, this study also found that whilst there is evidence of In-Group Advantage in the recognition of Joy, In-Group Advantage was not demonstrated for the other basic emotions under investigation. Tooby and Cosmides (1990) suggested an evolutionalry explanation for In-Group Advantage in the recognition of Happy, arguing that this emotion is important for cohesion and is less relevant to relations outside the social group.

Whilst Sauter et al. (2010) found that basic emotions (Joy/Amusement, Sad, Angry, Fearful, Disgusted and Surprised) were reliably identified cross-culturally across two highly unrelated cultures (British English and Himba), the same study found that positive emotions (achievement/triumph, relief, sensual pleasure) were not as reliably distinguished cross-culturally, as discussed in 3.9.3 above. This suggests that the vocal expression of positive emotions may be more culture-specific. One explanation given for this is that the experience and expression of Happy can aid in-group social cohesion. This is supported in previous research on emotion and the regulation of interpersonal relationships (Shiota, Keltner, Campos & Hertenstein, 2004).

Cross-cultural recognition of Happy may rely more on facial expression. This is supported by cross-cultural studies of recognition of facial expressions of basic emotions (e.g. Ekman & Friesen, 1971) which tend to find that Happy is the most reliably recognised emotion from facial expression. From an evolutionary perspective, facial expression would only be observable in close physical proximity and may also signal submission, as discussed in Chapter One.

Given that the smile causes adjustments in the vocal apparatus and tends to be auditorily perceptible, it is interesting that vocal expressions of Happy appear to be more difficult to recognise cross-culturally than Anger and Sadness. This aspect is also investigated in the present study.

One possible explanation suggested by Juslin and Laukka (2003), for the high degree of accuracy found in the cross-cultural decoding of Sad vocalisations, is that it may be easier to distinguish between high arousal and low arousal emotions and studies have often included several high arousal emotions (such as Angry, Fearful and Happy) and only one low arousal emotion (usually Sad) and Neutral which may also be regarded as low activation.

However, this argument does not explain why in previous studies Angry tends to be decoded with a greater level of accuracy cross-culturally than Happy or Fearful, all three emotions being classed as high arousal emotions. Since 1872, evolutionary psychology has offered the explanation that Angry may be more accurately decoded as from an evolutionary perspective this emotion would entail threat and potential risk to one's survival (Darwin, 1998).

Patterns of confusion are reported to be very similar cross-culturally in previous studies.

Van Bezooijen et al. (1984), for example, reported similar patterns of confusion cross-culturally between Dutch, Taiwanese and Japanese listeners rating vocal emotion expressed by Dutch speakers. Juslin and Laukka (2003) point out that previous studies rarely present results in the form of a confusion matrix. However, where this has been done, patterns of confusion between emotions have also been found to be similar cross-culturally. In particular, confusion seems to occur between emotions which have similar levels of arousal (e.g. Van Bezooijen et al., 1984 Pell et al., 2009a; Scherer et al., 2001). For example, confusion between Happy, Anger and Fearful has tended to be attributed to the high level of arousal of all of these emotions.

3.9.5 *Verbal and vocal cues*

Previous research (Ishii et al., 2003; Kotz & Paulmann, 2007; Kitayama & Ishii, 2002) suggests that there is possible cultural variation in the relative weight given to the use of vocal and verbal cues in the communication of emotion. In Kitayama and Ishii (2002), 29 Japanese and 29 American undergraduates (both males and females) listened to 360 words in both their original recorded form and in the content-filtered form and rated the pleasantness of the vocal tone of each utterance from 1 ("very unpleasant") to 7 ("very pleasant"). They found that vocal cues had a greater influence than verbal cues in the interpretation of emotion by Japanese speakers whilst native English speakers (American English speakers) seemed to depend more upon verbal cues than vocal cues. This would lead one to anticipate that Japanese listeners would generally decode emotion from solely vocal cues more reliably than American English listeners.

3.9.6 *Vowel quality*

Results in previous studies have occasionally hinted at the possible influence of vowel quality upon the expression and recognition of vocal emotion. The articulation of vowels in particular may be affected by smiling during a Happy vocalisation. Laver (1981) and Van Bezooijen (1984) presented conflicting evidence on the influence of vowel quality on the recognition of the vocal expression of emotion.

Van Bezooijen (1984, p.30) suggested that "extra... lip-rounding is easier to detect in rounded vowels, whereas extra lip spreading is easier to detect in unrounded vowels".

According to Van Bezooijen (1984) this means that happy vocalisations may be easier to detect on [i] due to the extra lip-spreading effect of smiling. She commented that Laver (1981) had suggested the reverse relationship between vowel quality and perception of lip-rounding, implying that Happy would be easier to detect when encoded on [u], given the distortion of this vowel which would be created by smiling.

Ohala (1984) argued that the technique of vocal tract lengthening, signalling a larger sound source is used by certain animals when expressing anger or aggression and that since [u] necessitates a lengthening of the vocal tract, it is more likely to be used sound symbolically to suggest aggression or anger.

Van Bezooijen (1984) suggested that to investigate further the possible effect of vowel quality on emotion recognition, vowels used should, as far as possible, remain constant between one phrase and another without sacrificing plausibility of utterances. However, Van Bezooijen (1984) also commented that controlling for vowel quality in semantically neutral sentences across different languages would be at best extremely difficult.

The mono-cultural study by Xu and Chuenwattanapranithi (2007) found that vowels synthesized with dynamically protruding lips were more frequently heard as Angry or spoken by a larger person and those with dynamically spreading lips were more frequently heard as Joyful or spoken by a smaller person (Xu & Chuenwattanapranithi 2007). This could potentially facilitate the recognition of vocalisations as Happy when vocalised on [i], for example, or possibly lead to over-interpretation of vocalisations as Happy, where other emotions are perhaps more likely to be decoded as Happy when they are vocalised on [i], for example.

3.9.7 *Utterance length*

Some previous studies suggest that recognisability of emotion increases with stimulus length. So, for example, it will be easier to recognise an emotion encoded on a full sentence than on a single word. For example, Erikson (2010) found that identification of emotion vocalised on single vowels was less accurate than identification on single words, again suggesting improved recognisability on longer utterances.

In the study by Beier and Zautra (1972), native speakers of American English vocalised

emotions on semantically neutral utterances of varying length. Decoding tests were performed by American English, Polish and Japanese listeners. This study found an In-Group Advantage for native (American English speaking) listeners in the shorter stimuli. However this In-Group Advantage diminished, the longer the stimulus, and no In-Group Advantage was evident in the longest stimuli which were full sentences.

Pell et al. (2009b p.432) commented that ”Controlling better for utterance complexity/length” remains “an ongoing challenge for researchers” (Pell et al. 2009b, p.432). Pell and Kotz (2011) investigated how quickly basic emotions were recognised from prosodic cues. They found that recognition accuracy for Happy required longer cues compared to Sad, Angry and Fearful.

Although related, utterance length is different from the acoustic measure of duration which is discussed in Chapter Two.

3.9.8 *Gender and age*

Cross-cultural studies of the vocal communication of emotion have not tended to test for the influence of gender upon encoding and decoding of vocal emotion. In any case, most studies do not have large enough participant groups to be able to make reliable comments upon such influence. However, initial studies (e.g. Audibert et al., 2008; Dromey et al., 2005; Erickson, 2006; Scherer et al., 2001) report that women have been found to be more accurate than men in decoding emotion, although little difference is sometimes found. For example, although Scherer et al. (2001) found that female listeners from nine different cultures identified vocal emotion significantly more accurately than male listeners overall, the difference between the two groups was slight - 67% for females and 65% for males.

Erickson (2006) found that for all three cultural groups studied, (Japanese, American English and Korean), female listeners identified each emotion more accurately than male listeners, except for Happy which American English males and females identified with the same level of accuracy, Suspicion which Korean males identified more accurately than Korean females and Angry which Korean males and females identified with the same level of accuracy. Audibert et al. (2008) found that although the raw data suggested that female listeners discriminated emotion more accurately and with a higher

confidence rating than male listeners, they found no statistically significant effect for listener gender.

Pell (2001) commented that there was evidence in his study to suggest that gender was significant in the prosodic encoding of emotion, since of the ten encoders (five female and five male), the four encoders whose vocalisations were most accurately recognised were all female (Pell, 2001). This was a mono-cultural study (Canadian English) focussing on Happy, Sad and Angry. Scherer et al. (2001) also found that there was a significant effect for actor gender although it was suggested that there were too few participants (two male and two female) to draw any conclusions from this.

Dromey et al. (2005) found that there was a small inverse correlation between age and decoding accuracy and that those participants who had learnt a second language seemed to recognise emotions more accurately than those who had not when decoding vocal emotion expressed by speakers from their own culture.

This Chapter has reviewed the methodologies and findings of previous cross-cultural research which has investigated how reliably emotion is recognised from vocal cues, both in-group and cross-culturally. Previous research on the characteristics of the vocal correlates of emotion has been discussed in Chapter Two. Findings and questions arising from previous studies helped to form the basis of the four research questions for this study, explained in Chapter Four.

Chapter Four

Research questions

4.1 Introduction

Chapters One, Two and Three have reviewed previous research to demonstrate the context and motivation for the present study, which aims to contribute to the understanding of the vocal correlates of emotion and the extent to which these are similar cross-culturally.

As discussed in detail in Chapters Two and Three, there is some previous evidence of beyond chance in-group and cross-cultural decoding of basic emotions from vocal cues. This indicates that there are significant distinctions between the vocal cues used to signal different emotions and that at least some of these cues are similar cross-culturally.

Whilst there is evidence in previous decoding experiments that in-group and cross-cultural decoders can distinguish basic emotions from vocal cues beyond chance level, further cross-cultural research is necessary to investigate whether this is a pan-cultural phenomenon. More research is also needed to investigate further the distinctive acoustic patterns of each basic emotion which allow vocally expressed emotion to be accurately decoded and to ascertain the extent of cross-cultural similarities and differences in these cues.

Whilst in previous mono-cultural and cross-cultural studies in this area, there is evidence of vocal correlates which appear to distinguish high arousal emotions from low arousal emotions, there is some conflicting evidence on this, especially in the few cross-cultural studies which have included analysis of vocal cues and knowledge of the vocal correlates which distinguish emotions beyond high and low arousal remains more elusive. This study investigates evidence of acoustic cues which distinguish emotions, also beyond the arousal dimension. The answer to these questions will provide some insight into the differential influence of culture and potentially pan-cultural, phylogenetic, psycho-physiological cues on the vocal communication of emotion.

There is evidence in previous research (3.9.4) that different basic emotions are recognised at different levels of accuracy and it has often been suggested that Angry and Sad tend to be the emotions which are most accurately recognised from vocal cues.

This study aims to investigate whether Angry and Sad are the most accurately decoded emotions from vocal cues both in-group and cross-culturally for English and Japanese. Results of acoustic analysis are examined for evidence of cross-cultural similarities and differences in vocal cues of each emotion. It has sometimes been suggested in previous research that Happy is more accurately decoded relative to other emotions in-group than cross-culturally and that this may be due to this emotion being evolutionarily useful for social cohesion. It was predicted that Calm may share some of the acoustic features of Sad (a low arousal emotion) with some of the acoustic features of Happy (a positive valence emotion).

Previous cross-cultural studies (3.9.2 and 3.9.3) have tended to find that decoders decode vocal cues of emotion more accurately where these cues are produced by decoders from their own, or a more similar cultural group. (Kitayama & Ishii, 2002; Menezes et al. 2010; Scherer et al., 2001; Thompson & Balkwill, 2006). However, there is conflicting evidence of In-Group Advantage and the Cultural Proximity Hypothesis.

Pell et al. (2009a) concluded that "…linguistic similarity is not a consistent factor which predicts the accuracy of vocal emotion recognition across languages." (2009a, p.117). American English decoders in the study by Erickson et al (2006) identified Japanese expressions of Happy, Sad, Angry, Surprised and Suspicious vocalised on the Japanese word "banana" more reliably than Japanese decoders. Except for Happy, Korean decoders also decoded all the emotions more reliably than the in-group Japanese decoders and identified Japanese Angry at a higher level of accuracy (100%) than either Japanese or American English decoders. The authors suggested that the in-group Japanese decoders may have been distracted from prosodic cues by "an anomaly between lexical meaning and prosodic meaning" (Erickson et al., 2006, p.4). Whilst "banana" is a loanword from English, Erickson et al. (2006) suggest that American English decoders may not have been distracted in the same way as in-group Japanese decoders as they may not have decoded the Japanese pronunciation of the loanword "banana".

Erickson (2010) found that Japanese and American English decoders identified emotion vocalisations by a Korean speaker more accurately than Korean decoders identified the emotions in the same vocalisations. The authors suggested that since the native Korean decoders knew that the vocalisations were produced by a Korean encoder, they may have been expecting lexical cues to the emotions and suggested that a future study may

present the data to Korean decoders without informing them that the vocalisations were Korean.

Himba speakers in the study by Sauter et al. (2010) were less accurate overall in their decoding of emotion from vocal signals such as laughs and screams than were native English English speakers whether within-group or cross-cultural decoding was observed. The authors suggested that this was probably due to the English group being more familiar with psychological testing than the Himba group.

Decoding results of the present study were examined for evidence of In-Group Advantage by comparing in-group and cross-cultural decoding accuracy by Japanese and English decoders. Results of acoustic analysis were also examined for evidence of cross-cultural differences in vocal cues of emotion which may explain any In-Group Advantage.

There have been hints in previous studies of the possible influence of vowel quality on the communication of emotion (Van Bezooijen, 1984; Laver, 1981). Laver (1981) suggested that Happy would be easier to detect when encoded on [u], given the distortion of this vowel which would be created by smiling. Van Bezooijen (1984) suggested the reverse relationship between vowel quality and perception of lip-rounding, commenting that the Happy vocalisations on the semantically neutral utterances containing [i] in her study may have been easier to detect due to the extra lip-spreading effect of smiling. Xu and Chuenwattanapranithi (2007) supports Van Bezooijen (1984), finding that those with dynamically spreading lips were more frequently heard as Happy.

A previous study by Ohala (1984) argued that vocal tract lengthening, signals a larger sound source and is used by certain animals when expressing Angry and that since [u] necessitates a lengthening of the vocal tract, [u] was more likely than other vowels to signal Angry. This suggestion is supported by Xu and Chuenwattanapranithi (2007) who found that vowels synthesized with dynamically protruding lips were more frequently heard as Angry or spoken by a larger person.

Van Bezooijen (1984) suggested that to investigate further the possible effect of vowel quality on emotion recognition, vowels used should, as far as possible, remain constant between one phrase and another without sacrificing plausibility of utterances but also commented that controlling for vowel quality in semantically neutral sentences across

different languages would be at best extremely difficult. The present study tested for potential vowel influence on the vocal communication of Happy, Sad, Angry, Fearful and Calm in English and Japanese by controlling for vowel quality in three pseudo-utterances which were phonotactically possible in both English and Japanese.

Research into a wider variety of cultures and further direct comparison between unrelated cultures will further understanding of these questions. Largely due to the methodological difficulties involved, there have been very few examples of balanced studies in this area. A key challenge in the present study was the construction of an appropriate methodology with the goal of answering these questions.

This balanced cross-cultural study of two cultures, differentiated by unrelated languages, was constructed to address these questions. The encoding and decoding experiments were designed to be symmetrical to facilitate cross-cultural comparison of both decoding results and vocal correlates.

In order to further understanding in this area, the following specific research questions were investigated in the present study. The first of these questions relates to in-group decoding of emotion. The second question addresses to cross-cultural decoding of emotion. Question three was constructed to test for In-Group Advantage and the fourth question tests for the influence of vowel quality on emotion decoding.

4.2 In-group decoding

To what extent can vocal cues signal emotion to people from the same cultural group, comparing English and Japanese production and perception of Happy, Sad, Angry, Fearful and Calm?

This question investigates English and Japanese production and perception of the emotions Happy, Sad Angry, Fearful and Calm to find whether the encoders produced distinctive vocal cues which signal these emotions to participants from the same cultural group. There is consideration of whether some emotions are easier to decode than others in group and whether there are patterns of confusion between emotions when decoding in-group which may help to explain why some emotions are well-decoded or poorly-decoded.

Acoustic analysis was conducted on the most reliably decoded data to investigate significant vocal cues of each emotion in each culture and for each culture there is consideration of the extent to which vocal cues found to distinguish emotions can help to explain any different decoding levels for different emotions and confusion patterns in the decoding of emotions.

There is cross-cultural comparison between in-group decoding results and vocal cues found to significantly distinguish emotions, with the goal of finding whether levels of in-group decoding vary cross-culturally and how this may be reflected in vocal cues.

Where any emotions are particularly well-decoded or poorly-decoded in-group and where there are strong patterns of confusion, these findings are compared cross-culturally.

4.3 Cross-cultural decoding

To what extent can vocal cues signal emotion to people from different cultural groups, comparing English and Japanese production and perception of Happy, Sad, Angry, Fearful and Calm?

Research question two investigates whether English and Japanese participants can recognise beyond chance emotions expressed by speakers of each other's culture and acoustic analysis tests how the answer to this question is reflected in cross-cultural similarities and differences between the vocal cues used to signal each emotion.

Patterns of confusion between emotions are also examined and the data is examined acoustically to investigate how patterns of confusion are reflected in the level of distinctiveness of vocal cues for these emotions.

4.4 In-Group Advantage

Are vocal cues of emotion more easily decoded by people from the same cultural group as the speaker, comparing English and Japanese production and perception of Happy, Sad, Angry, Fearful and Calm?

This question tests for evidence of In-Group Advantage when in-group and cross-cultural decoding levels are compared.

4.5 Influence of vowel quality

Can vowel quality influence reliability of decoding of Happy, Sad, Angry, Fearful or Calm by English and Japanese decoders?

This question is concerned with investigating the possible influence of vowel quality influences the decoding of emotion. Both in-group and cross-cultural decoding results are analysed for any effect of vowel quality. The vowels included in this study and the issue of how comparable vowels in English and Japanese are realised is discussed in Chapter Five (5.4).

These research questions were addressed by the construction of a balanced methodology with cross-culturally symmetrical encoding and decoding experiments (Chapter Five) and by acoustic analysis of the most reliably decoded data. Results of the decoding test are discussed in Chapter Six) and evidence from acoustic analysis is explained in Chapter Seven. Chapter Eight draws together the results of decoding tests and acoustic analysis with the aim of answering these research questions.

Chapter Five

Methodology of encoding and decoding experiments

> Developing reliable, valid, and effective techniques to elicit emotions in a laboratory setting remains a key challenge for researchers (Anolli et al. 2008, p.569).

5.1 Introduction

In order to test the hypotheses explained in Chapter Four, a symmetrical, simulation/self-induction method was constructed to collect data in which the Happy, Sad, Angry, Fearful and Calm were encoded in vocal expression and both native speakers of English and native speakers of Japanese were included as both encoders and decoders. Language difference was used as a proxy for cultural differentiation as discussed in 3.2. The participant profile is explained in 5.2.

Chapter Three described in detail the methodological challenges inherent in cross-cultural studies of the vocal expression of emotion. Consideration of a trade-off between authenticity and consistency of data needs to be made: simulated data is regarded as artificial whilst natural data from spontaneous speech has problems of consistency, especially in cross-cultural research. Where the aim is to investigate vocal cues of emotion, if emotions are encoded on verbal utterances, appropriate masking procedures are required to eliminate verbal cues: this was discussed in 3.8. Cross-language research into emotion vocalisation must also deal with the issue of translation, both of emotion words and of any verbal utterances: see 3.6.1.

As discussed in 3.5, there have been calls for 'balanced' and 'symmetrical' studies in this area (Galatà & Romito, 2010; Pell et al., 2009a; Pfitzinger et al., 2011; Scherer et al., 2001; Thompson & Balkwill 2006); however, the challenges inherent in constructing a symmetrical design, in which the encoding and decoding procedures are cross-culturally consistent, helps to explain the lack of such studies.

The present study is the first balanced cross-cultural study of the prosodic correlates of emotion to incorporate symmetrical experimental design, where the encoding and

decoding procedures are cross-culturally consistent. This consistency facilitated direct cross-cultural comparison of results of both decoding tests and acoustic analysis in order to test the hypotheses detailed in Chapter Four. This meant that considerable attention had to be paid to the development of an appropriate methodology before testing could take place in the present study. This chapter describes how this study addressed these issues in its experiments to encode and decode emotion. Chapter Seven explains the methodology for acoustic analysis of the data in this study.

The encoding experiment was constructed to produce data which was as segmentally consistent as possible for both groups to minimise segmental influence (apart from vowels) and to facilitate cross-cultural comparison in decoding accuracy and acoustic patterns associated with particular emotions. During the design phase of the experiments, native speaker input was elicited on which emotion labels to use and on the construction of pseudo-utterances which were also phonotactically possible in both Japanese and English.

All participants encoded all emotions investigated on three pseudo-utterances which were phonetically consistent and phonotactically possible in both English and Japanese and which differed from each other only in the vowel used (see 5.4). This gave cross-cultural consistency, allowed the isolation of vocal cues from verbal cues of emotion, avoided problems of translation and meant that vowel quality influence could be tested. Contradictory evidence of vowel influence on emotion decoding has been found in previous studies (Laver, 1981; Van Bezooijen, 1984). The methodology of the present study was also constructed to allow possible vowel influence to be highlighted.

A symmetrical design was facilitated by the use of a simulation/self-induction procedure to gather the data in the encoding experiment. As discussed in 3.7, the distinction between natural, induction and simulation procedures to collect data on vocal expression is not necessarily clear cut. Most previous studies of emotion vocalisation, including cross-cultural studies have used a simulation method, where emotions are acted, which, as discussed in Chapter Three, can sometimes lead to induction of the emotion vocalised, possibly by a feedback process (see Chapter One). As argued in 3.7, such expressions could be described as authentic expressions of emotion. Vocal expressions of emotion are also acted in the present study.

5.2 Participant profile

Four female Japanese informants helped in the choice of emotion labels to be used as emotion labels in Japanese. These informants were native speakers of Japanese, aged 22-35 and were fluent speakers of English, having also lived in the UK for at least three years. Each of these informants was completing or had just completed a course at Newcastle, Northumbria or Sunderland University.

Eight female native speakers of English (E) and eight female native speakers of Japanese (J) took part in the encoding experiment.

Four female native speakers of English (E) and four female native speakers of Japanese (J) rated the data in the reliability test. These were different from those who participated in the encoding experiment. This test was conducted prior to the decoding test to reduce the amount of data presented to the decoders.

Eight English (four female and four male) and eight Japanese (four female and four male) participants took part in the decoding test. All subjects were different from those who participated in the encoding experiment and the reliability test. Sample size was too small to test for gender influence. It might be argued that males have a harder task since they are decoding vocalisations by speakers of a different gender. However, it is not unreasonable to assume that males may be as attuned to in-group female expression as they are to male expression, and as equally challenged by the cross-cultural material as female listeners.

Each participant completed a profile questionnaire which can be found in Appendix A. The questionnaires showed that age range of participants was limited to 18-35 and all participants were from broadly the same social group, if indeed we can interpret the fact that participants were university students as being indicative of them being from the same social group. This therefore excluded the influence of age and social group upon the communication of emotion in the voice, since the groups were too small to consider these effects. All encoders, reliability test raters and decoders were students from Newcastle University and Northumbria University. The questionnaire showed that the English participants did not speak Japanese and had not travelled to any Far Eastern country. Japanese encoders, reliability test raters and decoders had studied English at school in Japan. However, Japanese participants (encoders, reliability test raters and decoders) had not previously travelled outside Japan and had only recently arrived in

the UK, within a month of taking part in the test. Japanese participants indicated that foreign films which they had watched were dubbed in Japanese.

One reason for choosing to conduct the experiments with native speakers of Japanese was that native speakers of Japanese fitting the participant profile could be recruited locally. Japanese was compared to English also because, as discussed in Chapter Three, English and Japanese are unrelated languages, which could act as a proxy for greater cultural distance in the search for evidence of the possible influence of covariant psycho-physiological influence on the vocal correlates of emotion. There has also been less research on Eastern cultures than on Western cultures in this field.

The influence of gender upon the encoding of emotion is not tested here. There were too few participants to provide sufficient comparison and normalisation would have been a further complication in acoustic analysis if male and female voices were to be compared. Female speakers were selected for the purposes of this study also because anecdotally, Japanese females are reputedly less likely than Japanese males to hide emotion and there is generally less research into the female voice. There is strong anecdotal evidence that Japanese males are generally much more reticent and inhibited in their expression of emotion due to emotional expression being stigmatised and seen as "weak" “feminine” and "childish".

5.3 Emotion labels

One complication for researchers setting up cross-cultural research in this area is the controversy which exists regarding the possible cultural bias attached to the emotion words used in these studies (3.6.1).This study investigated Happy, Sad, Angry and Fearful which, according to the psychology literature, are viewed as "basic" (1.3 and 3.6.1). The basic emotion labels used for English (Happy, Sad, Angry and Fearful) have been commonly used in previous research to refer to these emotions. ‘Calm’ was included for the reasons discussed below. An attempt was made to ameliorate the issue of translation by focussing particularly on affect bursts, remembered experiences and facial expressions of emotion and by using carefully chosen Japanese words, selected during the design phase of the experiment. These words were selected by careful consultation with four native speakers of Japanese, who were asked to suggest emotion

labels for the four basic emotions to be investigated and for 'Neutral' and 'Calm' for the reasons explained in 5.4.1 below.

Basic emotion	**Japanese**	**Affect bursts**	**English translations and further detail**
Happy	**ureshi -i**	(wow; yatta; lucky)	happy, ecstatic-e.g. reaction to good news (1st person)
Happy	**tanoshi-i**	wa (female only); ooo	enjoyment: feeling within a situation - having a good time (1st person)
Sad	**kanashi-i**	haaa - sigh, whisper	sad (Anecdotal evidence from the Japanese informants suggested that native speakers of Japanese were inhibited in the vocal and verbal expression of Sad and that Sad was more likely to be expressed by facial expression.
Sad	kawai-so	None suggested	sad (3rd person)
Sad	hikam	None suggested	hopeless; despairing; pessimistic (more formal, less common)
Sad	nage-ku	None suggested	grief (written)
Sad	yuu-utsu	None suggested	melancholic (1st person)
Angry	**atamani-kuru**	None suggested	anger - quite colloquial, 1st person possible
Angry	oko-ru	None suggested	anger (2nd/3rd person)
Angry	hala-ga-tetsu	saite - "you're the worst! I hate you!)	anger/annoyance (1st person) - it was suggested that a lower voice conveyed greater anger.
Angry	ika-ru	None suggested	anger (older form; formal; literary)
Fear	osoro-shi	None suggested	fear/trepidation - sounds "old-fashioned and "corny"
Fear	**koa-i**	None suggested	fear/trepidation (1st person) - commonly used.
Fear	koa-garu	None suggested	fear/trepidation (3rd person)
Fear	kyo-hu	None suggested	terror/panic
Calm	**shiawase-na**	mmm	peaceful contentment; joy
Calm	**ochitsuite**	None suggested	with equanimity, collectedly, with composure, self- possessed; calm' and as having no connotation of suppressing emotion. Viscerally unperturbed.
Calm	yorokobi	None suggested	bliss (written, very formal)

Table 5.1: Basic emotion labels. Emotion labels written in bold were regarded by the native Japanese speakers as fulfilling the following criteria: a) Must be expressible in the first person (necessary condition) b) Must be in oral use – not used in written form only (necessary condition) c) No other word is to be seen as more generic within any selection of possible labels which seem to fall within a similar category. d) Commonly used; informal.

Table 5.1 shows a summary of the results of discussions with four native speakers of Japanese during the design phase of the encoding experiment. English to Japanese translations are given and specific sounds which informants suggested they may produce if they were experiencing the particular emotion are shown in brackets.

There seems to be a lack of clear guidelines as to the linguistic criteria of 'basic' or 'primary' emotions. The following criteria were constructed in collaboration with the Japanese informants in an attempt to be more explicit than previous studies in how the labels were derived. The Japanese emotion labels used in this study therefore fulfilled the following criteria, which it could be argued can be considered important when describing an emotion as basic:

a) Must be expressible in the first person (necessary condition)

b) Must be in oral use – not used in written form only (necessary condition)

c) No other word is to be seen as more generic within any selection of possible labels which seem to fall within a similar category.

d) Commonly used; informal

The emotion labels written in bold were regarded by the native Japanese speakers as fulfilling the criteria a, b, c and d above. Of these, they agreed that Ureshi. Kanashi-i, Atamani-kuru, Koa-i and Ochitsuite were the most appropriate to use. For brevity, the English terms Happy, Sad, Angry, Fearful and Calm are generally used in this study to represent both English and Japanese labels unless specific reference is being made to the Japanese terms.

5.3.1 *Calm*

Other studies of emotion vocalisation have sometimes included 'neutral' as a category, which is sometimes used as a control base against which to measure the vocalisations of other emotions. It was originally also the intention of this study to include neutral as a category and native speakers of Japanese suggested 'mukando' as a translation of 'neutral'. However, they indicated that 'mukando' can be ambiguous in that it could suggest that the person is viscerally unperturbed in the sense of calm, which is seen as positive or perhaps more negatively that they are feeling indifferent, the latter being the

more likely interpretation according to the Japanese informants. Native speakers of English also suggested that ‘neutral’ could sometimes be interpreted as positive and sometimes as negative. Neutral/mukando was therefore viewed as ambiguous in terms of valence. It was decided to include a second positive emotion, Calm, in addition to Happy, instead of Neutral. Calm was viewed more as unambiguously positive and viscerally unperturbed.

Japanese informants suggested ‘ochitsuite’ as a translation of ‘calm’ and defined ‘ochitsuite as ‘with equanimity, collectedly, with composure, self-possessed; calm’ and as having no connotation of suppressing emotion. The Japanese and English native speakers who acted as informants indicated that ‘ochisuite’/‘calm’ was a clearer category and less open to different interpretations than mukando/neutral as well as expressing more accurately the idea of ‘viscerally unperturbed’ than ‘neutral’/’mukando’.

5.4 Pseudo-utterances

The pseudo-utterances were practised before the recorded game section until the encoder felt confident producing them. Encoders were asked to vocalise emotion on these pseudo-utterances in the game task, from which all the vocal expression data was gathered.

The use of pseudo-utterances has several advantages. They help to avoid the influence of the verbal channel without the need for masking techniques which may cause vocal channel distortion. Problems of translation are also avoided. See section 3.8.

In addition, pseudo-utterances were constructed which were as segmentally consistent as possible across the two languages to facilitate cross-cultural comparison and to highlight possible influence of vowel quality. The pseudo-utterances therefore needed to be phonologically viable and phonotactically feasible sequences in both languages. See section 2.2 for further discussion of the different rhythmical structures of English and Japanese.

This is a novel approach. The use of relatively consistent segments cross-culturally also meant there was less likelihood of segmental interference in acoustic analysis.

5.4.1 *Influence of vowel quality*

Conflicting suggestions regarding possible influence of vowel quality have been hinted at in previous studies (Laver et al. 1981; Van Bezooijen 1984). This is discussed in more detail in section 3.9.6. As discussed in Chapter One, Ohala (1984) also suggested that [i] and [u] are psycho-physiologically significant, and have phylogenetic influences.

The pseudo-utterances constructed for the present study were also used to highlight the influence of vowel quality and to compare any influence cross-culturally. Each pseudo-utterance contained only one vowel quality, which varied from utterance to utterance. Consonants were held constant for all three utterances. The three vowels tested included the most open vowel, the most front close vowel and the most back close vowel used in the language of each culture with the aim of testing the suggestion made in previous studies that vowel quality may influence communication of emotion. For English, the open vowel varied between [a] and [ɑ] depending on the pronunciation of the pseudo-utterance by each encoder. The open vowel in Japanese was the open centralised vowel [ä] which is somewhere between English [a] and [ɑ]. The close front vowel for English and Japanese was [i]. The close back vowel for English was [u]. The Japanese close back vowel is pronounced with neither fully rounded nor fully spread lips [ɯ]: the lips are compressed towards each other but are neither rounded like [u] nor spread like [ɯ].

Henceforth, the symbols [a]/[ɑ] refer to the open vowel, the close front vowel is symbolised by [i] and the close back vowel is symbolised by [u]/[ɯ]. It was anticipated that the differential spreading in [u] and [ɯ] may have some influence on any vowel quality effect. There was some variation in the realisation of all of these vowels. However, phonetic symbols are placed in square brackets as they are not phonemic representations.

The encoding experiment was also constructed to allow vowel influence to be investigated more systematically.

5.4.2 *The form of the pseudo-utterances*

The pseudo-utterances used were the same for both cultures to allow cross-cultural comparison of decoding of vocal cues of emotion without the possibility of segmental

influence, except for the possible influence of vowel quality (5.4.1). Using pseudo-utterances and varying the vowel as described meant that the verbal channel could be "masked" without interfering with the vocal channel. If any of the phrases prove to be deviant, the cause will evidently be the particular vowel quality, thus preventing the occurrence of unexplained deviant phrases, which can skew results as found in previous research as discussed in 3.9.6. As far as possible, vowels and consonants common to the two language systems were used, thus reducing the problem of 'strange sounding' sounds which may influence the interpretation by one language group of an emotion expressed by another group.

Native speakers of Japanese and native speakers of English confirmed that these utterances carried no verbal meaning in either language, were phonotactically possible in both languages (see 5.4.2) and were not suggestive of any particular emotion for either cultural group. Two sets of orthographic utterances were found, a shorter set and a longer set. Since more reliable emotion identification levels had been found for longer than for shorter utterances, it was decided to use only the longer utterances for the encoding test in this study. Testing for length of utterance influence on emotion decoding would have necessitated too much data in the decoding experiment. However, keeping the length of orthographic length of utterances constant meant that this factor did not influence results as has sometimes been the case in previous studies (3.9.7). The significance of duration on the other hand could still be tested.

Pseudo-utterances were of the same length in the sense that they each consisted of five syllables in a CV CV CV CV CV pattern. See 2.2 for discussion of the different rhythmical structures of English and Japanese and pitch accent in Japanese. The consonant pattern remained the same for each utterance. Only one phonological vowel quality was used in each utterance and this varied from utterance to utterance, allowing the influence of vowel quality upon the decoding of emotion to be tested as discussed in 5.4.1.

The use of two-syllable pseudo-words was considered; however, this was problematic for various reasons. It would have been more difficult to find pseudo-utterances which were phonetically and phonotactically possible in both languages which also carried no verbal meaning in either language and contained a consistent vowel in each utterance.

Words of more than one syllable would also lead to the need for English subjects to decide on which syllable carried the primary stress.

Since the vocalisation may have been influenced by whether the encoder interpreted the pseudo-utterance as a whole utterance or as isolated words, each pseudo-utterance was presented as a “nonsense-sentence”. Japanese informants in the design phase of the experiment also demonstrated that pitch accent varied depending upon whether the written symbols were seen as forming a sentence or several isolated words.

Each pseudo-utterance was presented on a separate card to the encoder. The following pseudo-utterances, written in roman script, were presented to the English encoders as “nonsense sentences”, which was made more evident by the use of capital letters and full-stops:

Cha na ga ma ba.

Chee nee gee mee bee.

Choo noo goo moo boo.

Four different writing scripts were considered for the written pseudo-utterances presented to Japanese encoders. Four native speakers of Japanese who acted as informants in the design phase of this experiment agreed that whilst it would be preferable to present the emotion labels in kanji script as this was the normal script used, it would be more appropriate to present the pseudo-utterances in the simpler and less common katakana script.

After input from native Japanese speakers involved in the design phase of the experiment the decision was made to present the pseudo-utterances to the Japanese encoders in katakana script. The Japanese informants indicated that katakana script is easier to read than kanji, hiragana or roman script. They commented that the more traditional hiragana script learnt first by children in schools is more complex and reportedly more difficult to read for native Japanese speakers, as are roman letters used phonetically. The informants also suggested that hiragana script may appear childish.

Again, it was explained to the Japanese encoders that the katakana symbols on each card constituted a "nonsense sentence" in order to avoid the production of each katakana symbol in isolation which could have affected the pitch accent used. This was already evident in the Japanese script since pauses or isolated words would have had to have been indicated by additional marks in the script.

5.4.4 *Isolating vocal cues from verbal cues*

In the present study, the inclusion of encoder groups who spoke different languages was used as a means of accessing speakers from different cultures. This is not to say that all native speakers of the same language necessarily have the same cultural identity. Indeed, what is classed as the English language includes many diverse varieties, expressing a very broad variety of cultural identities. The fact that the American English speakers in the study by Kitayama and Ishii (2002) were more influenced by verbal than vocal cues did not mean that this would also be the case for the British English speakers in the present study. The few previous studies which have looked at the interplay between vocal and verbal cues are discussed in Chapter Three and are referred to in 6.7 and the small number of studies which have investigated the connection between vocal and facial cues are discussed in section 3.7.5.

This study does not test for the relative weight given to vocal and verbal cues by English and Japanese encoders. However, if either Japanese or English encoders were found to decode emotion from solely vocal cues less accurately than the other encoder group this could suggest greater reliance on verbal, facial or body cues, although this is not necessarily the case. The relevance of this for the present study is discussed in section 6.7.

5.5 Optional stimuli for self-induction of emotion

It could be argued that simulated data is not authentic in the sense that it is not spontaneous data in an ecologically valid setting. It is possible that because of this, the data was more devoid of cultural influence. It is also possible that the laboratory setting may have different connotations for different cultures; indeed individual reaction to the laboratory setting may also have varied.

The distinctions regarding authenticity drawn between spontaneous data gathered from a normal setting, simulated emotion and induced emotion are not necessarily clear-cut (Scherer, 2013b). In the encoding experiment for the present study, authenticity in the sense of participants experiencing the emotion they were expressing was encouraged.

The procedure aimed to stimulate self-induction as a means to encouraging authenticity as suggested by Scherer (2013b), who argued that there is not a clear dividing line between simulation and induction of emotion, the former potentially leading to the latter. Whilst some examples of vocalisations derived from natural settings may not represent authentic emotions, it is also possible that some examples of simulated emotions were experienced by the speaker during their vocalisation. The blurred distinction this creates is relevant to discussions comparing techniques for collection of data on emotion.

The method used in the present study therefore combined simulation and self-induction. Whilst the encoders simulated the emotions, various optional techniques were suggested to them to help them produce these emotions and all participants reported using one or more of these techniques and reported feeing one or more of the emotions they produced in the experiment. It is recognised that self-reporting cannot be assumed to be reliable; however, as Cowie and Cornelius (2003) point out, it is difficult to see how we might access an encoder's feeling without asking them.

As part of the warm-up activity, participants in the encoding experiment were also asked to remember a time when they experienced the emotion suggested to them by each facial expression in turn. Remembering these experiences provided an optional stimulus in the game section of the encoding experiment. This relates to the influence of memory in the experience of emotion discussed above.

Pictures of quasi-universally decoded facial expressions were used, in addition to the carefully selected lexical labels for the emotions. It should be noted that this is a study of the vocal communication of emotion and the quasi-universally decoded pictures of facial expressions of Happy, Sad, Angry and Fearful were used only to help address translation issues and to act as one of several optional methods to focus encoder attention on the emotions to be encoded, possibly also encouraging self-induction of the emotions. The pictures of facial expressions tested by Ekman and Friesen (1975) were of Caucasian faces, whilst those tested by Matsumoto (1996) were of Japanese faces. Since all encoders were female and half the encoders were English and half were Japanese, for each encoder, it was decided to use one picture of a female face for each emotion from the Caucasian faces in Ekman and Friesen (1975) and one picture of a female face for each emotion from the Japanese faces in Matsumoto (1996). The pictures are not reproduced here for copyright reasons.

The design of this study also aimed to address the issue of possible cultural bias in the emotions studied and labels attached to these emotions. Photographs of facial expressions of basic emotions were used, which had been found to be pan-culturally recognised. Both Caucasian (Ekman and Friesen, 1975) and Japanese (Matsumoto, 1996) photographs were used for both English and Japanese participants. During the design phase of this study, both Caucasian (Ekman and Friesen, 1975) and Japanese (Matsumoto, 1996) photographs were used in interviews with native English and native Japanese speakers in order to elicit consensus on appropriate English and Japanese terms to use as descriptors for these expressions. Both sets of photographs and the language-specific verbal labels were then used in the encoding and decoding experiments since it was considered that these would allow greater focus and clarity and the labels themselves may be an effective stimulus in the encoding experiment.

It was suggested to the encoders that they may wish to focus on the facial expression corresponding to the particular emotion before they vocalised the emotion and whilst vocalising the pseudo-utterances and to imitate the expression if they wished. Emotion could possibly have been induced due to emotional contagion (see 1.5.2) from focussing on the facial expression of an emotion or from a facial feedback mechanism (see 1.5.1) if encoders intentionally or unintentionally imitated the facial expression of an emotion.

Since there was no tested pan-culturally recognised facial expression for Calm available, it was suggested to encoders that to help them to produce the Calm vocal expressions, they may wish instead to think of a piece of music they knew which made them feel Calm and remember how they felt when listening to this. The encoders heard music when they entered the room to participate in the encoding experiment. It was intended that this would provide an example of a piece of music which induced Calm. When asked if the music they heard made them feel any differently, all participants said that it made them feel ‘relaxed’ or ‘calm’. Self-reporting is not necessarily reliable, although as pointed out by Cowie and Cornelius (2003), it is difficult to see how we might access an encoder’s feeling without asking them.

The music played was in a major key and was slow. As discussed in 3.7.2, there is empirical evidence that music in a major key can express and induce emotions which have a positive valence (Webster & Weir, 2005; van der Zwaag et al., 2009), whilst

slow music can express and induce emotions which are classed as low arousal (Pinchot-Kastner & Crowder, 1990; Webster & Weir, 2005). Unlike many experiments which test specifically for musical induction, encoders were not asked to focus on feeling Calm whilst listening to the music. In addition, even if Calm was induced by this music, this cannot be taken as evidence that this induction was simply due to the music being slow and in a major key since the music has many other attributes, which may have been influential.

The music used was western music, though from previous eras and could perhaps be considered to be biased towards the English encoders. However, there is evidence that there is sensitivity to emotion in music from a culture which is not one's own (Balkwill & Thompson, 1999). Considerable time elapsed between the participant hearing the music and their participation in the game task from which the data was gathered, so there is no suggestion made here that the vocal expressions of Calm were induced by the music the encoders heard when initially entering the room.

One of the novel aspects of the present study is that affect bursts were used to focus the encoders' attention on the emotions and as a possible means of self-induction. It has been claimed (Scherer, 1994; Krumhuber & Scherer, 2011) that affect bursts may be primitive forerunners of expression of basic emotions (see Chapter One for further discussion). The encoder was asked to think of affect bursts which they may associate with each of the emotions in the warm-up activity and then to remember these as an optional stimulus in the final section of the experiment from which the encoded data was gathered.

Imagining an experience during which they would feel the emotion was suggested to encoders as an optional aid to help them to focus on the emotion they were aiming to vocalise. This technique may also have caused self-induction of emotions.

5.6 Encoding experiment

5.6.1 *Pre-test*

All participants were informed that the aim of the research was to investigate the expression of emotion in the voice. They participated in a pre-test with the aim of preparing them for the encoding experiment. The pre-test questionnaires can be found in

Appendix B. This had the aims of focussing on the basic emotions which were to be vocalised in the individual interviews and introducing elements related to techniques for self-induction of emotion: encoders were asked to suggest any affect bursts they may associate with each emotion; they focussed on facial expressions representing Happy, Sad, Angry and Fearful (Ekman & Friesen, 1975; Matsumoto, 1996). The idea was to focus attention upon emotions to be considered via a visual stimulus which was common to both cultural groups, rather than primarily via verbal cues which would have to be translated between groups. Attention was therefore focussed on affect bursts and facial expressions and on encoders remembering how they felt or imagining how they would feel during experiences associated with emotions expressed in these facial expressions and associated affect bursts.

There was no visual (facial expression) stimulus for Calm as there was no evidence of this in research into universal facial expression of emotion. In the second section of the pre-test (Appendix B), encoders were asked to describe any experience in which they may experience Calm. They were also asked to suggest any sounds they or another native speaker of their language variety may make if they were feeling each emotion.

These techniques are akin to method-acting techniques used by actors to self-induce emotional states. This prepared encoders for the final section of the experiment from which the data was gathered (section 5.5). The techniques had the purpose of helping encoders to self-induce the emotions they were to express, although there is no claim made that the data from the encoding experiment were authentic expressions of emotion.

5.6.2 *Recording*

The encoded data for this study was elicited during the 'game' section of the encoding procedure was recorded. All recordings took place in the recording studio in the School of Education, Communication and Language at Newcastle University. Recordings were made on a Tascam DAT recorder (model DA-30) using an audio-technica omnidirectional condenser microphone (model ATM10a) mounted on a stand and boom, placed in front of the encoder at a distance of approximately 50cm from the encoder. It was decided not to use a lapel radio microphone since, given the nature of the task, encoder movement may cause sound interference. The microphone was placed

on a stand and boom rather than on the table to avoid the encoder accidentally knocking the microphone during the tasks.

Using this set-up, the potential for varying mouth-microphone distance will have influenced intensity measures. The problem of mouth-microphone distance is arguably particularly the case in experiments where the subject is attempting to express emotion, which may be likely to be accompanied by gesture and movement. Even with a lapel microphone, mouth-microphone distance would have varied, especially because of possible increased movement due to emotional expression. In addition, gestural movement would have been more likely to interfere with a lapel mic or headphones. Whilst the use of a uni-directional headset microphone may have resulted in less variation in mouth-microphone distance, it was decided that wearing a headset may cause the subject to be more inhibited in their attempts to express emotion and may interfere with natural movement which could help with their expression of emotion and the production of more ecologically valid vocalisations with more valid f0 measurements. It should be recognised, however, that this method will have influenced intensity measures, so results regarding intensity are tentative and further evidence of any findings will be required in future research.

5.6.3 *Encoding experiment procedure*

Encoders were interviewed individually. The whole procedure took approximately 25 minutes for each encoder. The following is a summary of the procedure to encode vocal expressions of Happy, Sad, Angry, Fearful and Calm.

The interviewer used English emotion labels with the English participants and Japanese emotion labels with the Japanese participants.

The following gives a brief outline of the procedure of the encoding experiment. Further details on specific elements are given in the sections below.

- The three pseudo-utterances were introduced as “nonsense sentences” and encoders were informed that later they would be acting these with different emotions. The encoder practised the pseudo-utterances until they felt confident saying them.

- Various techniques were suggested to the encoders to help them produce the pseudo-utterances with each emotion, including recalling any sounds (affect bursts) they might associate with each emotion, concentrating on the facial expression pictures for a particular emotion and remembering or imagining a situation or experience during which the encoder felt or would feel the emotion. Encoders were also free to use any other technique they felt may be useful. They were informed that there was no facial expression picture for Calm, but that thinking of a piece of music which made them feel Calm and remembering how they felt might help when saying the "nonsense sentences" in a Calm way.
- Recorded 'game' task from which the encoded data was gathered.
- Feedback.

The encoding test was conducted by the same experimenter for both cultures in an attempt to gain consistency in how the experiment was conducted. The experimenter was a native speaker of English. The Japanese encoders did not appear to have a problem understanding the English instructions and the game section of the experiment from which all of the encoding data was gathered was conducted in the encoder's own native language, However, it should be noted that the Japanese encoders may have been influenced by having to interact with an experimenter who was not a native speaker of Japanese.

5.7 Game task

The game task was conducted after the warm up tasks. All of the encoded data for this study was collected from the game section. The first vocalisation of each emotion on each pseudo-utterance with which the participant was satisfied was edited out to be used in a reliability test (see 5.9).

The experiment was conducted in the encoder's native language. The experimenter responded with English emotion labels for the English participants and with Japanese emotion labels for the Japanese participants. Interaction in this section of the test was minimal and limited to vocalisation of the emotions by the encoder, the encoder stating immediately after a vocalisation if they were not satisfied with their expression, the

experimenter guessing which emotion had been vocalised and the encoder confirming whether or not the experimenter had guessed correctly.

The encoder was informed that the next part of the test was a game. They were asked to choose one of the five emotions and to vocalise this on the first pseudo-utterance. They were informed that the instructor would try to guess the emotion they vocalised whilst her back was turned so that judgements would need to be made from the voice alone. As discussed in Chapter Three, Friesen (1972) and Ekman (1972), also reported in Matsumoto (2006), found evidence that Japanese participants tended to mask emotion when not alone. The interviewer turned her back in the encoding procedure used in the present study in an attempt to reduce the masking of facial expression during vocalisation by both the Japanese and English participants. It is of course possible that one culture may be more likely than another to mask their vocal expression of emotion.

Participants were asked to inform the interviewer straight after a vocalisation if they were not satisfied with the vocalisation and wished to repeat it. In these cases, the participant repeated the vocalisation until they were satisfied. Only when a participant was satisfied with their vocalisation, the interviewer took a guess at which emotion the participant had intended to express. In order to access more examples for each emotion, when the experimenter felt that she had guessed correctly on the encoder's first attempt at expressing the emotion, she gave an alternative answer in order to elicit further examples. The encoder then repeated the task. It should be noted that even when the experimenter thought that she had guessed correctly, this may or may not have been the case. In repeating the task, it is possible that the encoder produced a clearer example of the expression of the emotion. However, for consistency and to help to reduce experimenter influence, whether or not the interviewer guessed the emotion correctly, the first vocalisation of each emotion on each pseudo-utterance with which the participant was satisfied was then edited out to be used in the reliability test (5.9).

For each pseudo-utterance, this procedure was repeated for each emotion. Encoders were informed that they could repeat emotions; otherwise the last one for each pseudo-utterance would be obvious to guess. At least two examples of each emotion were elicited for a pseudo-utterance before moving on to the next pseudo-utterance.

The game aimed to incentivise the encoders to produce vocal expressions of emotions which were recognisable and to disinhibit encoders by playing a game and focussing attention on the instructor who guessed which emotions were vocalised by the encoder.

The fact that the instructor turned her back so that the encoders' possible attempts to make facial expressions of emotion could not be seen also meant that the encoder could rely only on vocal cues to communicate the emotion on the nonsense utterances. If the encoder chose to use the optional stimuli these may have resulted in self-induction of the emotions.

5.8 Post-experiment feedback

There is recognition in the present study that emotions may be self-induced by portraying them. Various optional stimuli were also suggested to the encoding participants with the aim of aiding emotion vocalisation and encouraging self-induction of emotions. There was no attempt to record whether participants experienced the emotions they vocalised during the encoding procedure itself so there is no claim made that any specific vocalisation of emotion in the data set expressed an authentic experience of the emotion. Future studies may perhaps record when emotion is induced by, for example, the encoding participant pressing a button for a specific emotion.

Encoders were, however, asked for informal feedback on the game task after the encoding experiment. They were asked which emotions they found easier or more difficult to vocalise, whether they had experienced any of the emotions, which of the optional techniques they used and which they found most helpful. No statistical analysis was conducted on this feedback. All of the encoders reported feeling one or more of the emotions they vocalised at least some of the time as they were vocalising them and one encoder claimed that she had experienced all of the emotions, although not necessarily during all vocalisations. This constituted a form of self-induction, possibly similar to method-acting, as described above. However, which emotions were reported to have been experienced varied from one encoder to another. Which emotions encoders found easier to vocalise also varied, as did the techniques encoders chose to use and which they found most helpful.

5.9 Reliability test

In the encoding experiment, each speaker (eight E and eight J) was asked to vocalise three pseudo-utterances, each containing a different vowel quality ([i], [u] and [a]).

Each of these utterances was vocalised with five different emotions (Happy, Sad, Angry, Fearful and Calm). The data sample from the encoding tests thus consisted of 120 utterances for each culture, making a total for both cultures of 240 utterances:

2 cultures x 8 speakers x 3 utterances x 5 emotions = 240 utterances

This was too large a sample to present in the decoding test and it was decided to reduce the data for presentation to decoders. If the decoders were to listen to all of the encoded data, this would have necessitated their attending tests on three occasions. This would have made it more difficult to recruit decoders and it was quite possible that participants may only complete part of the decoding experiment. Presenting the data at a single sitting also facilitated attendance of Japanese decoders within a month of their arrival in the UK.

A reliability test was therefore conducted, with the aim of selecting encoders from each culture who most convincingly encoded emotion from the perspective of in-group decoders. A copy of this test can be found in Appendix C. The data from these three speakers for each culture was then used in the decoding test in order to limit the amount of data presented to decoders.

Four raters for each culture performed the reliability test, which was conducted in the recording studio in the School of Education, Communication and Language at Newcastle University.

In order to avoid the same problems for the raters as the decoders associated with the size of data presented to them, the decision was made to limit the amount of data presented to raters in the reliability test. Only in-group data encoded on [tʃa na ga ma ba] was used. Each test therefore included a total of 80 utterances:

1 culture x 8 speakers x 1 utterance x 5 emotions = 80 utterances

Whilst it is acknowledged that including the data on all three vowels may have yielded different results, since the data was to be reduced, the decision to use only the vocal expressions encoded on [a] rather than [i] or [u] was made due to bearing in mind the potential influence on emotion expression of the latter two vowels, as discussed above.

Raters for each culture performed a forced-judgement test for each utterance to indicate which emotion they thought was expressed. They also gave a rating on a scale of one to five of how convincing they felt the emotional expression was. The three highest

scoring encoding participants each scored an average of at least three out of five for reliably decoded emotions overall by each of the in-group raters. A reduced data sample of 45 utterances by the English encoders and 45 utterances by the Japanese encoders was therefore selected in reliability tests performed by raters of the same culture as the encoders. The total sample size for the decoding test was then:

2 cultures x 3 speakers x 3 utterances x 5 emotions = 90 utterances

5.10 Data sample for the decoding experiment

The reduced data sample (90 utterances) selected in the reliability test was randomly ordered and presented to participants in the decoding experiment. These were preceded by six practice utterances to accustom the participants to the task they were being asked to perform. These six practice utterances were randomly selected from the utterances which had been excluded by the reliability test.

Sequentially ordered numbers preceded each utterance. Each utterance was repeated three times with a precise two-second interval between each and an eight second interval before the following number.

5.11 Decoding experiment – location, equipment and procedure

The decoding experiment took place in the public auditorium (The Herschel Building) at Newcastle University.

The participants in the decoding experiment entered the auditorium and sat far enough away from each other for the questionnaires of other participants not to be legible to them. Participants were asked not to interact with other people in the room. After a brief welcome, the task was then explained to the decoders.

It was explained that each person would be given a handout to complete whilst they listened to some recordings. For each item they heard, they were asked to choose one emotion word from a choice of five and not to leave any answers blank. Emotion words were in the participant's native language and were the same words as those included in the encoding experiment. The rationale for choosing these particular words is given 5.5.

Decoders were also asked to give a confidence rating on a scale of one to three for each judgement they made. It was anticipated that these confidence ratings could be useful in interpreting the relative salience of acoustic correlates between one realisation and another. In the end, it was decided that this was beyond the scope of the present study and these confidence ratings were not used further.

The recording was played through a computer and loudspeakers. It should be noted that if decoders had heard the recordings through headphones, the vocal cues of emotion may have been clearer to the decoders, which may have resulted in increased percentage ratings for recognition of emotions. It could also be argued that presenting the recordings via loudspeaker may encompass more the concept of redundancy of vocal cues present in normal speech, discussed earlier. It is suggested that in spontaneous speech, there tends to be redundancy of cues in the vocal signal to allow for interference in the transmission phase of communication. The forced judgement decoding test which the participants completed can be found in Appendix D.

The next chapter discusses the results of the decoding test.

Chapter Six

Decoding test results

> While there is an abundance of studies of how facial expressions of emotion are communicated and recognised in a cross-cultural setting…, surprisingly little data have been gathered on how vocal expressions of emotion are recognised by individuals from different linguistic and cultural backgrounds (Pell & Skorup 2008, p.519).

6.1 Introduction

The decoding test was set up to investigate the research questions posed in Chapter Four, which were based on evidence and questions arising from previous research discussed in Chapter Three. The methodology used to gather the recognition data and participant profiles were explained in Chapter Five. There have been calls in previous research for balanced cross-cultural studies. However, methodological difficulties of this approach have led to few examples of balanced studies in this area as discussed in Chapter Three. This cross-cultural study is unusual in that it is a balanced study, with the aim of comparing in-group and cross-cultural recognition of emotion from vocal cues encoded by each cultural group.

In addition, this is the first study to use symmetrical cross-cultural experimental procedures and the aim of this approach was to facilitate direct cross-cultural comparisons between in-group and cross-cultural decoding results. The methodology for the encoding and decoding experiments of the present study is explained in Chapter Five.

The data sample collected from the decoding experiment is given in section 6.2 below. A confusion matrix illustrating mean percentage in-group and cross-cultural identification of emotion according to emotion target by English and Japanese decoders can be found in Table 6.1.

The results of the decoding test are discussed, with reference to their relevance for the research questions and hypotheses presented in Chapter Four. The data which were found to be most reliably decoded were analysed acoustically as explained in Chapter Seven. Acoustic analysis investigated evidence of vocal cues which signal significant

distinctions between emotions and between the English and Japanese vocal expressions of these emotions.

6.2 Data sample

The reduced data sample of 90 utterances which was selected in the reliability test (5.10) was presented to participants in the decoding experiment. The profile of participants was explained in 5.2. Decoding results presented here are based on 8 decoders from each culture. Both male and female decoders participated in the decoding experiment. All participants were different from those who participated in the encoding experiment. A total of 1440 utterances from the decoding test were obtained:

2 cultures x 8 decoders x 90 utterances = 1440 responses

The items from the decoding test which were found to most reliably communicate emotion according to the emotion target were analysed acoustically. The data sample for acoustic analysis is explained in 6.11.

6.3 Better-than-chance ratings

Previous studies, as also reported in meta-analyses (Elfenbein, & Ambady, 2002; Juslin & Laukka, 2003) and reviews (e.g. Scherer, 2003), often use forced judgement tests to investigate decoding accuracy. Where mean decoding accuracy for a group of decoders is found to be 'better-than-chance', this is taken as evidence that a decoder has identified an emotion according to target. Whilst chance levels are not precise because chance can be due to any number of reasons, ratings tend to be well beyond chance levels and are generally considered to be a useful tool to allow comparison of results across studies. This study uses the same method and chance level was taken as 20% as there were five possible options in the forced judgement test.

6.4 In-group emotion recognition

Overall, English decoders decoded English vocal expressions of emotion according to the emotion target beyond chance (68%) and Japanese decoders also decoded Japanese vocal expressions of emotion according to the emotion target beyond chance (49%), demonstrating that the pseudo-utterances contained vocal cues which signalled emotion to decoders from the same cultural group.

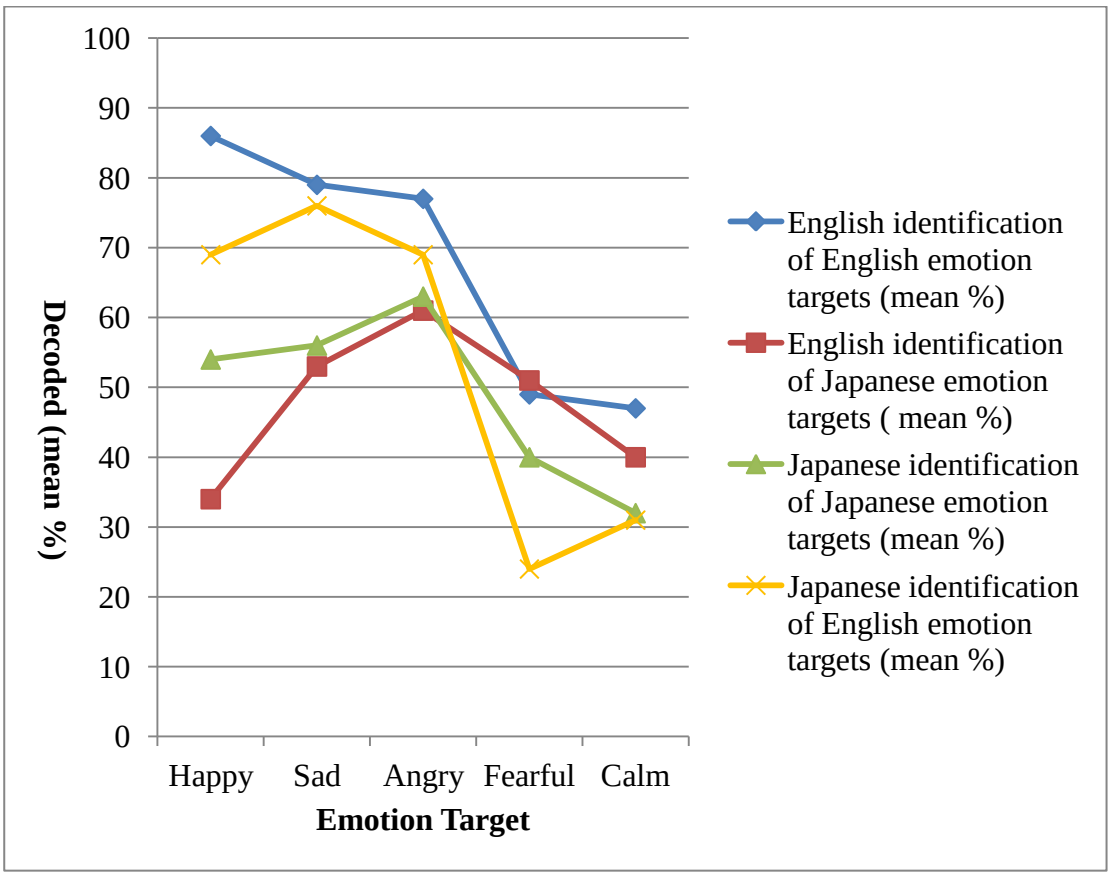

Figure 6.1: The lines represent mean percentage in-group and cross-cultural identification of each emotion target by English and Japanese decoders.

Some emotions were decoded at higher levels than other emotions as illustrated in Figure 6.1 above, which presents the mean percentage identification of each emotion target by decoders. Both in-group and cross-cultural levels are shown.

The graph in Figure 6.1 illustrates relative in-group and cross-cultural identification of each emotion target by English and Japanese decoders.

Happy, Sad and Angry were decoded well above chance level by both cultures when decoding vocal expressions of emotion encoded by their own cultural group. This

suggested that significantly distinct vocal correlates of these emotions were present in the vocal expressions of these emotions for each culture. Acoustic analysis was conducted to search for these vocal cues (Chapter Seven). See 6.5 for discussion of whether there is evidence that each culture identified similar vocal cues as communicating Happy, similar vocal cues as communicating Sad and similar vocal cues as communicating Angry. As can be seen in Figure 6.1, levels of in-group identification according to emotion target for Happy, Sad and Angry were much lower for Japanese, at 54%, 56% and 63% respectively, compared to English at (86%), (79%) and (77%) respectively. This aspect is discussed further in 6.6.

Culture	**Emotion target**	**Mean % identification of emotion targets by English decoders**					**Mean % identification of emotion targets by Japanese decoders**				
		Happy	Sad	Angry	Fearful	Calm	Happy	Sad	Angry	Fearful	Calm
English	Happy	**86**	4	3	1	6	**69**	8	11	4	7
	Sad	2	**79**	0	6	13	3	**76**	3	8	10
	Angry	4	6	**77**	1	10	10	4	**69**	8	8
	Fearful	6	26	7	**49**	12	6	49	15	**24**	7
	Calm	12	38	2	2	**47**	24	40	3	3	**31**
Japanese	Happy	**34**	2	30	11	23	**54**	4	15	7	19
	Sad	0	**53**	0	26	21	1	**56**	10	10	24
	Angry	15	0	**61**	5	19	17	4	**63**	3	14
	Fearful	13	4	15	**51**	18	13	6	32	**40**	10
	Calm	4	33	0	23	**40**	10	43	8	7	**32**

Table 6.1: Confusion matrix illustrating mean percentage in-group and cross-cultural identification of emotion according to emotion target by English and Japanese decoders. Percentages are rounded to the nearest whole number.

A confusion matrix showing in-group and cross-cultural decoding levels can be found in Table 6.1 above. Fearful and Calm were more poorly-recognised in-group than Happy, Sad and Angry by both English and Japanese decoders. English Fearful (49%) and Calm (47%) and Japanese Fearful (40%) were identified in-group beyond chance. In-group decoding of Japanese Calm (32%) was decoded beyond chance, although at a lower level beyond chance. High levels of identification of both Happy and Angry

indicate that despite these emotions both falling within the same level of arousal, they tended not to be confused with each other. Acoustic analysis was carried out to investigate what the distinct vocal cues for Happy and Angry were for each culture (Chapter Seven). See 6.5 for discussion of evidence to suggest whether or not there are cross-cultural similarities between the vocal cues of each of these emotions.

Japanese Calm was the only emotion to be identified in-group as another emotion more often than it was identified as its target, being identified as Sad (43%) more often than it was identified as Calm (32%). English Calm (47%) was also often identified in-group as Sad (38%). This suggests that vocal correlates of Calm are similar to those of Sad for both cultures, supporting previous evidence of confusion between emotions of similar level of arousal. Acoustic analysis was conducted to search for these vocal cues (Chapter Seven). See 6.5 for discussion of whether there is evidence that each culture identified similar vocal cues as communicating Calm and identified similar vocal cues as communicating Sad.

In-group, where English Fearful (49%) was identified as a different emotion, it tended to be decoded most often identified as Sad (26%). Japanese decoders in-group, identified Fearful as Angry almost as often as they decoded Japanese Fearful according to target.

Japanese Sad (56%) was also sometimes identified in-group as Calm (24%), although this occurred at only slightly beyond chance level. Where English Sad (79%) was identified as a different emotion, this was most often Calm (13%), although this was at well below chance level. This suggests that the vocal cues for Sad are more distinct than those for Calm for each culture. This observation was tested by statistical analysis of the results of acoustic analysis (Chapter Seven).

6.5 Cross-cultural emotion recognition

Overall, each culture identified emotion targets encoded by the other cultural group at better-than-chance levels, English decoders identifying 48% of Japanese emotion targets and Japanese decoders decoding 54% of English emotion targets. This supports findings of beyond chance cross-cultural decoding of emotion found in previous studies, demonstrating that at least some of the pseudo-utterances contained vocal cues which signalled emotion to decoders from the other cultural group.

Independently of possible vowel influence, cross-cultural and in-group decoding levels were above chance for each emotion except for Japanese decoding of English Fearful (24%) which was not considered to be decoded at a level reliably above chance.

Some emotions were decoded more reliably than others cross-culturally by both English and Japanese decoders. In common with in-group decoders, cross-cultural decoders demonstrated decoding levels well beyond chance for both English and Japanese Sad and Angry. Japanese decoders decoded English Sad and Angry at levels well beyond chance at 76% and 69% respectively, providing evidence of cross-culturally recognisable vocal cues in the English data for Sad and Angry. English decoders decoded Japanese Sad and Angry at levels well beyond chance at 53% and 61% respectively, also providing evidence of cross-culturally recognisable vocal cues in the Japanese data for Sad and Angry.

In common with English in-group decoders, Japanese decoders also decoded English Happy well above chance (69%), indicating that there were cross-culturally recognisable vocal cues in the vocal expressions of English Happy. However, English decoders identified Japanese Happy as Angry (30%) almost as often as they decoded Japanese Happy according to target (34%). Previous studies have also found confusion between decoding of high arousal emotions. See also 6.9 for a discussion of possible vowel influence on the decoding of Japanese Happy. In-group decoders decoded Japanese Happy at a much higher level (54%) than English decoders. This provides evidence of possibly culture-specific vocal cues of Japanese Happy.

Japanese decoders, like English in-group decoders, demonstrated lower decoding levels for English Fearful and Calm than for English Happy, Sad and Angry, decoding English Calm at 31% and decoding English Fearful at a level of 24%, which was at a borderline level with what would have been expected by chance and therefore not considered to be reliably above chance. English decoders of English Fearful tended to confuse English Fearful (49%) with Sad (26%) more than with any of the other emotions; Japanese decoders demonstrated an even stronger tendency to identify English Fearful with Sad, being twice as likely to identify English Fearful as Sad (49%) rather than Fearful (24%).

Beyond chance recognition according to target of Japanese Fearful and Japanese Calm by English decoders, at 51% and 40% respectively, also provided evidence of cross-culturally recognisable vocal cues.

Above chance cross-cultural decoding reliability indicated that common acoustic cues were used to communicate most of the emotions investigated in both cultures. Acoustic analysis was conducted (Chapter Seven) to search for evidence of cues common to both cultures.

6.6 In-Group Advantage and 'Cultural Cue Awareness' (CCA)

English decoders decoded 68% of English expressions according to the emotion target and 48% of Japanese expressions according to the emotion target. Japanese decoders decoded 54% of English expressions according to emotion target and 49% of Japanese expressions according to emotion target. The mean overall decoding reliability overall for both groups of decoders was 55%.

Previous cross-cultural studies have tended to find an 'In-Group Advantage' (3.9.2) in the decoding of vocal expressions of emotions, as well as support for the 'Cultural Proximity Hypothesis' (3.9.3). The present study used a balanced and symmetrical methodology (Chapter Five) to investigate whether English and Japanese participants demonstrated evidence of In-Group Advantage. Acoustic analysis was used to investigate cross-cultural differences in vocal cues of emotion, which would help explain differences in-group and cross-cultural decoding rates.

Evidence was found of In-Group Advantage for the English decoders, as English expressions of emotion were decoded significantly more reliably overall by English decoders (68%) than by Japanese decoders (54%), $F(1,88) = 4.070$, ($p = 0.047$). However, the difference between Japanese decoding levels (49%) and English decoding levels (48%) of Japanese expressions of emotion was found not to be statistically significant, $F(1,88) = 0.054$, ($p = 0.817$).

Closer examination of the data reveals that In-Group Advantage is emotion-dependent for the decoding of the English vocal expressions. English decoders demonstrated an In-Group Advantage compared to Japanese decoders when decoding English Happy, Fearful and Calm. English decoders decoded these emotions at mean percentage levels of 86%, 49% and 47% respectively, compared to Japanese mean percentage decoding levels of 69%, 24% and 31% for Happy, Fearful and Calm respectively.

Japanese decoders demonstrated an In-Group Advantage compared to English decoders when decoding Japanese Happy, which is decoded at a level of 54% by Japanese encoders and 34% by English decoders. Apart from identification of Japanese Happy, there is no evidence of In-Group Advantage in the decoding of the Japanese data. On the contrary, there is evidence that English decoders decoded Japanese Fearful and Japanese Calm at least as reliably as Japanese decoders decoded these emotions.

No In-Group Advantage was found in the decoding of English or Japanese Sad or Angry emotion targets. English decoding levels for English Sad (79%) and English Angry (77%) were similar to Japanese decoding levels for English Sad (76%) and English Angry (69%), and Japanese decoding levels for Japanese Sad (56%) and Japanese Angry (63%) were similar to English decoding levels for Japanese Sad (53%) and Japanese Angry (61%).

In-Group Advantage could be due to the influence of different display rules in each culture, including possibly different levels of restraint in each culture on the open expression of emotion. As discussed in Chapter Three, previous studies (Anolli et al., 2008; Soto et al., 2005) have found evidence of less distinct acoustic cues of emotion for East Asian cultures than for other cultures.

There is therefore some evidence that emotions are recognised more reliably than others both in-group and cross-culturally, although In-Group Advantage appears to be both emotion-dependent and culture-dependent. Since only two cultures were investigated, it was not possible to test for further evidence of the 'Cultural Proximity Hypothesis'.

The related concept of 'Cultural Cue Awareness' (CCA) is introduced in this study to encompass the idea that the cultural cues of emotion can be acquired or consciously learnt by members of different cultures, however close or distant their cultures are considered to be. Once more comprehensive details of vocal cues of emotion in different cultures have been found, it will be possible to test for evidence of CCA, focussing on recognition of vocal cues as well as cues in different modes and how these interact.

6.7 Psycho-Physiological Cue Awareness (PPCA)

It was found (6.6) that Japanese decoders did not demonstrate In-Group Advantage in decoding the emotions expressed vocally by Japanese encoders, except for Japanese Happy which was decoded by Japanese decoders at a level of 54% and by English decoders at a level of 34%. English decoders decoded Japanese Happy less reliably than they decoded any other Japanese emotion target. This evidence suggests that there could be cultural influence upon the acoustic correlates signalling Happy in Japanese.

However, Japanese decoders decoded English Happy (69%) more reliably than they decoded Japanese Happy (54%). Japanese decoders also decoded English Sad (76%) more reliably than they decoded Japanese Sad (56%), suggesting that the Japanese decoders were recognising vocal cues of Happy and Sad embedded in the English pseudo-utterances which they may not use themselves.

This suggests that the Japanese decoders were capable of recognising vocal cues of emotion which they do not use themselves, providing tentative evidence for the existence of acoustic cues of Happy and Sad in the English expressions which may be quasi-universally recognised even if, possibly due the influence of cultural display rules, they are not necessarily universally expressed, perhaps due to different levels of restraint related to emotional expression in the display rules of different cultures. Even though a decoder may restrain the expression of these cues themselves due to cultural display rules, they may well still be aware of these cues. This effect may be termed 'Psycho-Physiological Cue Awareness' (PPCA). In theory, the concept of PPCA may apply not only to the voice but also to other communication modes which may be psycho-physiologically influenced such as facial expression and body gesture.

An alternative or possibly additional explanation for Japanese decoders' high decoding accuracy of English emotion targets could be that the Japanese participants may have been more aware of the vocal correlates of emotion used by native speakers of English than originally anticipated, possibly due to exposure through the global mass media.

Another possible explanation for English expressions being more reliably decoded both in-group and cross-culturally than Japanese expressions may be that the English and Japanese participants may have been influenced differently by the experimental procedure itself. For example, Japanese encoders may have been more reticent than English encoders to produce the emotions in a laboratory setting.

Conversely, it is possible that cultural cues may be particularly prevalent for a specific culture, as could be evidenced by high in-group decoding levels and low decoding levels found consistently for cross-cultural decoders of these cues from other cultural groups. The concept of CCA is relevant here to encompass the idea that these cues could potentially be acquired or learnt, however close or distant the culture of the decoder is from that of the encoder. Again this concept may be applied to cues in other non-verbal modes of communication, such as facial expression and gestural cues.

Where cultural cues vary from one culture to another, the term 'Cultural Cue Awareness' (CCA) is also used to imply the possibility that people from different cultures can develop an awareness of each other's cultural cues of emotion.

The cues resulting from greater or lesser salience of psycho-physiological cues are also in a sense, cultural cues since salience could vary from culture to culture. It does not appear to be the case that the Japanese vocal expressions in the present study include strong cultural cues to the emotions expressed vocally since they are decoded with around the same level of accuracy by both Japanese and English decoders. It could be that emotions in Japanese are expressed more through verbal, facial or gestural modes than through the vocal channel, or if restraint is used in all modes in Japanese, it is possible that the use of silence may communicate emotion. This is a speculative observation; however it could merit further investigation.

Studies by Ishii et al. (2003), Kitayama and Ishii (2002) and Kotz and Paulmann (2007) suggested that different weights may be given to verbal and vocal information in different languages. The present study does not test for the relative weight given to vocal and verbal cues. However, the fact that English participants tended to decode Happy, Sad and Angry from solely vocal cues more reliably than Japanese participants, lends tentative support to the idea that the English group in this study may rely more upon vocal than verbal cues than do the Japanese participants. It may be the case that clearer vocal cues of emotion may be produced by the English group if they are more reliant on vocal than verbal cues.

It could be that emotions in Japanese are expressed more through verbal, facial or gestural modes than through the vocal channel and that native speakers of British English possibly rely more upon vocal than verbal cues than do the Japanese participants. However, Kitayama and Ishii (2002) found that native American English speakers were more dependent upon verbal than vocal cues when recognizing emotion

and Japanese speakers relied more on vocal cues than verbal cues. If we are to gain more insight into this area, it will be necessary to differentiate between different varieties of the same language since different varieties will tend to reflect different cultures, which will possibly use distinct patterns of cues in the communication of emotion.

It is also possible that for English, emotion may be distinguished as reliably or more reliably by verbal cues as by vocal cues. This would give a kind of belt and braces effect, constituting redundancy in the system of communication which would allow for interference in the transmission of the signal. This would be what would be predicted by Scherer's adaptation of The Brunswikian Lens Model (Scherer, 2003) if it is applied to the communication of emotion through a combination of modes (vocal, verbal and gestural).

Acoustic analysis was conducted to investigate why English Happy and Sad were more reliably decoded than the Japanese expressions of these emotions, including by Japanese decoders.

Japanese decoders decoded both Japanese Angry (63%) and English Angry (69%) at similar levels of reliability, which were both well beyond chance. Japanese decoders also showed almost identical, although fairly low levels of identification for Japanese Calm (32%) and English Calm (31%). Fearful was the only emotion which Japanese decoders decoded more reliably from the expressions encoded by Japanese encoders (40%) than by English decoders (24%). In addition, English decoders decoded Japanese Fearful at least as reliably as they decoded English Fearful.

English vocalisations of Sad, Angry and Happy are much more reliably decoded than Fearful or Calm by both groups. Japanese vocalisations of Sad, Angry and Happy are also more reliably decoded than Fearful or Calm by Japanese decoders. However, although English participants decoded Japanese vocalisations of Sad and Angry with higher percentage accuracy than Japanese Fearful or Calm, of the five emotions vocalised by Japanese speakers, English participants were least accurate in their decoding of Japanese Happy: see 6.9 for details of possible vowel influence. It is possible that there is a greater cultural influence upon the vocalisation of Happy in Japanese culture. The acoustic characteristics of the emotions studied are discussed further in Chapter Seven.

One explanation may be that English Happy vocalisations could be influenced by psycho-physiological response mechanisms which Japanese participants were able to decode despite possibly either not using these acoustic signals themselves or using them in a less pronounced way. This therefore constitutes possible evidence for the existence of 'Vocal Psycho-physiological Cue Awareness' for English Happy and acoustic analysis of this data may shed more light upon this suggestion.

Japanese decoding results for English Calm and English Fearful do not follow the same pattern. Whilst English Calm was decoded more reliably by English decoders, Japanese Calm was decoded at around the same level by both cultures. English Fearful was decoded at around the same level by both English and Japanese decoders. Japanese Fearful was the only emotion which Japanese decoders clearly decoded more reliably than English decoders.

There were similar reliability rates for Japanese decoding of Japanese and English Calm and Japanese decoders recognised Japanese vocalisations of Fearful more reliably than they recognised the English vocalisations of Fearful.

English decoders demonstrated that they were able to decode Japanese Fearful (51%) at least as reliably as they decoded English Fearful (49%). English decoders also decoded English Calm (47%) and Japanese Calm (40%) at similar levels.

6.8 Summary of patterns of confusion

The comments in this section are based on the confusion matrix in Table 6.2. Where confusion levels with another emotion were below 20% the specific percentage is indicated. Where an emotion target is identified as another emotion at between 20-30%, the emotion target is described as being sometimes confused with the other emotion. Where an emotion target is identified as another emotion at 30% or more, the emotion target is described as being often confused with the other emotion.

Japanese Happy was often confused with Angry and sometimes confused with Calm by English decoders. Japanese decoders were also more likely to confuse Japanese Happy with Calm or Angry, although identification with these emotions was low at 19% and 15% respectively. Confusion of Happy with both Calm and Angry could be explained by Happy being both a high arousal emotion, as is Angry, as well as a positive emotion,

as is Calm. Happy was more reliably identified than Calm, except for English decoding of Japanese Happy and Calm.

Japanese Sad was sometimes confused with Calm by both English and Japanese decoders. This could have been due to both Calm and Sad being low arousal emotions. However, Japanese Sad was also sometimes confused with Fearful by English decoders, which may have been due to both Sad and Fearful having a negative valence. English Sad was confused with Calm more than with any other emotion, although identification with Sad was only 13%.

Where English decoders confused Japanese Angry with another emotion, this was most likely to be Calm (19%) or Happy (15%). Where Japanese decoders confused Japanese Angry with another emotion, this was also most likely to be Happy (17%) or Calm (14%). Whilst confusion of Japanese Angry with Happy by both cultures may perhaps be explained by both these emotions being high arousal emotions, confusion of Japanese Angry with Calm by both cultures may perhaps be explained by some Japanese vocal expressions of Angry being particularly unconvincing or may be evidence of a restraint shown in the expression of Angry in the Japanese culture. Nevertheless, it should be remembered that Japanese Angry was decoded well beyond chance by both Japanese and English decoders, at 63% and 61% respectively.

English Fearful was sometimes confused with Sad by English decoders and Japanese decoders were twice as likely to decode English Fearful as Sad. Whilst Fearful and Sad are opposites on the arousal dimension, confusion between these two emotions could be due to both emotions having a negative valence. Japanese Fearful was often identified as Angry by Japanese decoders. This confusion could have arisen due to both Fearful and Angry being high arousal emotions.

Both English and Japanese Calm were recognised above chance both in-group and cross-culturally. However, Calm was often confused with Sad, the other low arousal emotion in the set, and Japanese decoders were more likely to decode both English and Japanese Calm as Sad. Japanese decoders also confused English Calm with the other positive valence emotion, Happy, at a level just beyond chance. Unexpectedly, English decoders sometimes confused Japanese Calm with Fearful, although this occurred at only just beyond chance level.

6.9 Vowel quality

The study was constructed in a way which allowed vowel influence upon emotion expression and decoding accuracy to be tested. In the encoding experiment emotions were encoded on three pseudo-utterances which vary segmentally only in the vowel quality used. Mean percentage in-group and cross-cultural confusion in identification of emotion targets on each vowel is illustrated in Table 6.2 below. The results in this table and in Table 6.3 below are discussed in more detail in relation to each emotion in sections 6.9.1-6.9.5.

Emotion Target	Vowel	Mean identification of emotion target (%)			
		English decoding English	English decoding Japanese	Japanese decoding Japanese	Japanese decoding English
Happy	overall	86	34	54	69
Happy	a	83	29	83	67
Happy	i	94	63	46	79
Happy	u/ ɯ	81	10	33	63
Sad	overall	79	53	56	76
Sad	a	85	50	42	58
Sad	i	85	48	63	96
Sad	u/ ɯ	67	63	63	75
Angry	overall	77	61	63	69
Angry	a	79	63	63	71
Angry	i	69	52	50	71
Angry	u/ ɯ	83	69	75	67
Fearful	overall	49	51	40	24
Fearful	a	60	50	50	21
Fearful	i	69	48	33	33
Fearful	u/ ɯ	19	54	38	17
Calm	overall	47	40	32	31
Calm	a	44	42	33	46
Calm	i	50	48	33	29
Calm	u/ ɯ	48	29	29	17

Table 6.2: Mean percentage in-group and cross-cultural identification of emotion by English and Japanese decoders, according to emotion target and pseudo-utterance vowel. Percentages are rounded to the nearest whole number.

Where there is common vowel quality influence found for both cultural groups, this lends support to possible universal influence although clearly further studies would be required in this area to lead to any stronger conclusions.

Some emotion-targets tended to be identified with other emotions at levels beyond chance (20%). Observation of decoding results for each vowel quality reveals possible vowel influence on these patterns of confusion.

Table 6.3 shows where emotions were identified as an emotion other than the target emotion at levels beyond chance, illustrating results according to vowel quality. Details in parenthesis give mean percentage identification with an emotion other than the target emotion. Confusion well beyond chance is highlighted in bold.

Vowel	Emotion Target	Emotion Targets confused beyond chance			
		English decoding of English	Japanese decoding of English	Japanese decoding of Japanese	English decoding of Japanese
a	Happy				**29 (Angry 48)**
a	Sad			**42 (Calm 33)**	**50 (Fearful 31)**
a	Angry				63 (Calm 23)
a	Fearful		**21 (Sad 63)**	50 (Angry 21)	50 (Angry 29)
a	Calm	**44 (Sad 52)**	**46 (Sad 46)**	**33 (Sad 33)**	42 (Sad 27; Fearful 23)
i	Happy				63 (Angry 20)
i	Sad				48 (Sad 27; Fearful 27)
i	Angry	69 (Calm 20)		50 (Happy 25)	52 (Happy 29)
i	Fearful		**33 (Sad 42)**	**33 (Angry 42)**	
i	Calm	50 (Happy 27)	**29 (Happy 33; Sad 33)**	33 (Sad 29)	**48 (Fearful 31)**
u/ɯ	Happy		63 (Sad 21)	33 Calm 29; Angry 25)	**10 (Calm 40; Fearful 23; Angry 20)**
u/ɯ	Sad	67 (Calm 25)		63 (Calm 21)	
u/ɯ	Angry		67 (Happy 21)		
u/ɯ	Fearful	**19 (Sad 52)**	**17 (Sad 42)**	**38 (Angry 33)**	54 (Calm 25)
u/ɯ	Calm	**48 (Sad 44)**	**17 (Happy 33; Sad 42)**	**29 (Sad 67)**	**29 (Sad 56)**

Table 6.3: Mean percentage in-group and cross-cultural confusion in identification of emotion targets on each vowel where confusion is beyond chance. Details in brackets show mean percentage identification with an emotion other than the target emotion. Confusion well beyond chance is highlighted in bold. Percentages are rounded to the nearest whole number.

Strong patterns appear to emerge here and comments below are based on instances of better-than-chance mean percentage identification with an emotion other than the target emotion. Comments are generally based on much better than chance ratings.

6.9.1 *Happy*

Table 6.2 shows that English Happy was most reliably decoded on [i] by both English and Japanese decoders, at 94% and 79% respectively, thus lending support to the suggestion in previous studies that Happy could be identified more reliably on [i], which has a shortened vocal tract and spread lips, as in a smile. See 3.9.6 for further discussion.

Japanese Happy was also most reliably decoded on [i] (63%) by English decoders, compared to [a] (29%) and [ɯ] (10%). However, Japanese in-group recognition of Happy on [i] does not follow what appeared to be an emerging pattern. Results also show that although Japanese Happy on [i] and on [ɯ] are both articulated with unrounded lips, Japanese in-group decoding of Happy was much more reliably decoded on [a], at 83% than on [i] (46%) or on [ɯ] (33%).

As can be seen from Table 6.3, English decoders were far more likely to decode Japanese Happy on [ɯ] as Calm (40%); identification according to target on [i] was 63%, whilst on [ɯ] target recognition was at only 10%, which is so far below chance as to be considered potentially relevant. It could be that it is not, or not only, spread lips suggestive of a smile which influences recognition of Happy according to vowel quality. Perhaps the back quality of [ɯ] influenced its recognition negatively. It could be that that the front quality of [i] is particularly relevant. Testing whether or not this is this case is beyond the scope of this study. However, it poses a question which merits further investigation in future studies. It is also possible that factors other than vowel quality, such as fundamental frequency, intensity and speech rate cues are at play here, influencing recognition of Japanese Happy. This is explored in the next chapter.

6.9.2 *Sad*

The strongest evidence found for the influence of vowel quality on the recognition of Sad was the Japanese decoders' high rate of recognition of English Sad on [i] (96%) compared to other vowels, although recognition was well beyond chance for all three vowels. As Table 6.2 shows, where Japanese decoders decoded English Fearful and English Calm as a different emotion from the target, there was a strong tendency to hear the vocal expression as Sad and Table 6.3 illustrates that this occurred independent of vowel quality. This is discussed further in 6.9.4.

6.9.3 *Angry*

In-group, Angry was most reliably decoded on the close back vowel, at 83% for English and 75% for Japanese. The evidence for the English expressions supports the suggestion (Ohala, 1984; Xu & Chuenwattanapranithi, 2007) that [u] is more likely to signal aggression. This was based on the idea that the vocal tract lengthening for [u] would be more likely to signal a larger sound source and phylogenetically would sound more threatening. This was discussed in 3.9.6. However, since the Japanese close back vowel, [ɯ], is unrounded, the evidence of better recognition on this vowel than on the other two does not lend support to this suggestion, although there may be some tentative support in the lower in-group recognition level for Japanese Angry on [i], at 50%, than on either other vowel, given that [i] is articulated with spread lips, which shortens the vocal tract.

6.9.4 *Fearful*

English decoders identified English Fearful at a level of 49%. However, English Fearful appears to be particularly difficult to communicate in-group on the vowel [u] compared to [a] and [i]. The decoding level on [u] (19%) did not reach chance level whilst English decoding rates were well beyond chance when English decoders decoded English Fearful on [a] (60%) and [i] (69%). This evidence lends support to the idea (3.9.6) that the vocal tract lengthening required for the production of [u] could signal a larger sound source and is more likely to sound Angry rather than Fearful; English Fearful tended to be heard in-group on [u] as Sad (52%).

Table 6.3 illustrates that the strong tendency for Japanese decoders to hear the English vocal expressions of Fearful as Sad occurred independent of vowel quality. No conclusions can be drawn regarding vowel quality influence for Japanese decoding of the English expressions, although it should be noted that the tendency for Japanese decoders to recognise English Fearful as Sad was strongest on [a] and [u]; these expressions were not recognised reliably beyond chance and were much more often recognised as Sad, at 63% and 42% respectively.

6.9.5 *Calm*

In-group, English decoders had a strong tendency to confuse Calm with Sad on [a] and on [u]. However, they most often confused English Calm with Happy on [i], which supports the suggestion that [i] is more likely to be heard as Happy. In conflict with this is the finding that Japanese decoders had a tendency to confuse English Calm with both Sad and Happy.

Japanese decoders identified English Calm according to target much more reliably on [a] than on [i] or [u], although this pattern is not replicated for Japanese or English in-group recognition of Calm, or for English identification of Japanese Calm expressions.

6.10 Summary

Decoding evidence suggests that there are vocal correlates which distinguish each emotion from the others for decoders from the same cultural group for both English and Japanese. There is also evidence that cross-cultural decoders can identify each emotion according to the emotion-target from vocal cues, with the possible exception of Japanese Fearful, which was not identified cross-culturally at a level which was reliably beyond chance. Confusions between emotions tended to follow dimensional lines of arousal or valence, except for decoding of English Fearful, which was often heard as Sad by in-group decoders and Japanese decoders were twice as likely to hear English Fearful as Sad.

Supporting evidence in previous studies (3.9.4), Sad and Angry were identified well beyond chance both in-group and cross-culturally. English Happy was also identified beyond chance in-group and cross-culturally. However, English decoders demonstrated a much lower level of reliability when decoding Japanese Happy, although the level was still considered to be beyond chance, at 34%. Unexpectedly, Japanese decoders decoded English Happy and Sad much more reliably than they decoded Japanese Happy and Sad, suggesting evidence of a newly observed phenomenon, coined here as ‘Psycho-Physiological Cue Awareness’ (PPCA).

Fearful and Calm were generally less well decoded by both English and Japanese decoders, in-group and cross-culturally, although English decoders decoded Japanese Fearful at least as reliably as they decoded English Fearful and English Fearful was the

only emotion which Japanese decoders decoded less reliably from the English vocal expressions than from the Japanese vocal expressions. Other than results for English Fearful, the Japanese vocal expressions were decoded at generally lower levels than the English vocal expressions by both in-group and cross-cultural decoders.

Evidence of In-Group Advantage in decoding was found for English decoding of Happy, Fearful and Calm and for Japanese decoding of only Happy.

Evidence was found of high in-group and cross-cultural recognition levels of English Happy expressed on [i] and English decoders recognised Japanese Happy more reliably when expressed on [i]. Where English decoders confused English Calm on [i] with another emotion, they tended to decode Calm as Happy, whilst the tendency was to decode Calm as Sad on the other two vowel qualities. This evidence supports the suggestion (Van Bezooijen, 1984) that Happy is easier to recognise on [i], due to spread lips, as in a smile. However, evidence found for Japanese in-group decoding appears to conflict with this. In addition, both in-group and cross-cultural decoding of Japanese Happy on [ɯ], which is also an unrounded vowel, was particularly low. It is suggested here that perhaps vowel influence on the signalling of Happy could be due to the front quality of [i] rather than or as well as its unrounded quality.

Angry was more likely to be decoded on the close back vowel, providing evidence in support of the suggestion by Ohala (1984) and Xu & Chuenwattanapranithi, (2007) that this vowel is more likely to signal aggression (3.9.6). However, conflicting evidence for the unrounded Japanese vowel [ɯ] requires further investigation in future studies. English in-group decoding of Fearful also supports this idea since a much lower decoding level was found for [u] (19%) than for [i] (69%) or [a] (60%). Based on these results, further cross-cultural research is merited on vowel quality.

Acoustic analysis was conducted to search for further evidence of the patterns found in decoding test evidence. There was a search for vocal cues which significantly distinguished each emotion in each culture, and significant vocal cues associated with each emotion were compared cross-culturally. Vocal cues were also tested for evidence of more significant distinction in the English expressions than in the Japanese expressions, in order to help explain the apparently greater salience in vocal cues in English than in Japanese for specific emotions which was evidenced in the decoding test. The results of acoustic analysis are presented in Chapter Seven with some reference to implications for decoding test results. Chapter Eight discusses patterns

which have emerged in the results, comparing evidence from the decoding test and acoustic analysis and draws together the conclusions of this study.

Chapter Seven

Acoustic analysis results

7.1 Introduction

Chapter Six discussed evidence of in-group and cross-cultural recognition of Happy, Sad, Angry, Fearful and Calm from vocal cues. The most reliably decoded data was analysed acoustically to investigate the vocal cues which signalled these distinctions. This Chapter presents the results of acoustic analysis and Chapter Eight discusses the results of acoustic analysis in relation to the decoding test results. The statistical analysis was carried out using the Statistical Package for the Social Sciences (SPSS). The aim was to highlight whether there was significant association between any acoustic patterning and emotion and culture.

In this chapter, summaries of results can be found in table and graph form and these figures form the basis of the discussion of results of the present chapter. The current data facilitated a more direct comparison by the use of data which was more phonetically consistent across pseudo-utterances and languages than that generally found in previous studies, both English and Japanese participants having vocalised the five emotions on the same three pseudo-utterances.

The sections below discuss the vocal cues of each emotion for each culture, followed by a cross-cultural comparison of the vocal cues of emotion with reference to findings in the decoding test.

7.2 Data sample for acoustic analysis

The decoding experiment tested the extent to which the encoders from each cultural group had embedded vocal cues of each emotion onto the pseudo-utterances. Utterances which were identified with the highest decoding scores were selected for acoustic analysis. The three vocal expressions of each emotion by each culture which were most reliably decoded by decoders from each culture were analysed acoustically. Vowel qualities were therefore not equally represented in the data which was analysed acoustically.

Where cross-culturally the three most reliably decoded pseudo-utterances were not the same, the size of the set was increased. For example, if items 3, 10 and 38 in the decoding test were those which were most reliably identified according to target for English Happy by English decoders, and items 10, 38 and 63 were those which were most reliably identified according to target for English Happy by Japanese decoders, the tokens included for acoustic analysis were 3, 10, 38 and 63. A total of 21 pseudo-utterances were selected from the English data and a total of 21 pseudo-utterances were selected from the Japanese data. Whilst the set of data included for acoustic analysis is a fairly small set, all data were reliably decoded by in-group decoders at levels well beyond chance. Selecting the most reliably decoded data meant that vowel qualities were not equally represented in the data which was analysed acoustically. When testing for significance of results, statistical tests were therefore also included to allow for inequality of group size.

As discussed in Chapter Three, the use of a decoding test to confirm the reliability of identification with the emotion target follows Scherer's adaptation of The Brunswikian Lens Model (Scherer, 2003). This model emphasises the importance of including both production and perception in the investigation of cues of communication. This model also suggests an element of redundancy in that not all possible cues are necessarily used all of the time, which allows for the possibility that some cues may be lost in the transition phase of communication.

Whilst previous studies of acoustic parameters of emotion vocalisation have found distinguishing features of high and low arousal emotions, finding acoustic features which may distinguish between discrete emotions, even basic emotions has proved more elusive. These measures are included here with the aim of contributing to the search for acoustic parameters which may be relevant to emotion discrimination. Acoustic analyses were performed on each of the selected pseudo-utterances. In total, eleven acoustic parameters were measured.

In the present study, all acoustic measurements were computed automatically by Praat speech analysis software (Boersma & Weenink, 2007). Pseudo-utterance start points and end points were decided according to the criteria detailed in 7.2.1 below. Acoustic measurements were then computed by Praat for each pseudo-utterance. A total of eleven acoustic parameters relating to duration, fundamental frequency and intensity were measured. These included Speech rate in milliseconds (SpRate), f0 in Hertz

(Maxf0), minimum f0 in Hertz (Minf0), range of f0 in Hertz (Rangef0), Meanf0 in Hertz (Meanf0), standard deviation of f0 in Hertz (SDf0), 90th percentile of f0 in Hertz (90thPercf0), 10th percentile f0 in Hertz (10thPercf0), 80 percent range of f0 in Hertz (80PercRangef0), mean intensity in decibels (MeanInt) range of intensity in decibels (RangeInt).

Statistical analysis of the acoustic measurements was conducted to determine significant vocal cues for the expression of each emotion in each culture. These emotion cues were then compared cross-culturally.

Each of the 42 pseudo-utterances was analysed acoustically and average results across speakers from the same culture are shown in below, alongside discussion comparing results for the English and Japanese data.

7.3 Speech rate

Each of the pseudo-utterances in the present study contained a CVCVCVCVCV sequence. In the design phase of the experiments, Japanese informants confirmed that the pseudo-utterances each contained five 'syllables' or 'morae'. For brevity, in the present study, the word 'syllable' is used to refer to each of the five CV units in the pseudo-utterances. The structure of the pseudo-utterances is explained in more detail in 5.5. Whilst five syllables constitutes a fairly short utterance, this is not unusually short for studies in this area which investigate SpRate. For example, in Erikson et al. (2008a) the word 'banana' was used as a stimulus. Despite being only three syllables long, significant differences in SpRate were found for culture and emotion in the study by Erikson et al. (2008a).

It is possible that in-group variation may have occurred in the vocalisation and decoding of emotion in this study due to regional influences upon both English and Japanese encoding and decoding participants. The sample sizes were too small to test for this. It is argued (Warner & Arai, 2001) that pitch accent does not affect durational measurements. It is beyond the scope of the present study to test for the possible influence of pitch accent upon results for f0.

In the present study, as in Pell et al. (2009b), speech rate in syllables per second was measured to investigate its significance in distinguishing between different emotions for

each culture. The pseudo-utterances were also tested for any main effect for culture. Speech rate was calculated by dividing the number of syllables in a pseudo-utterance by the duration (in seconds) of each pseudo-utterance. f0 track peaks were observed and it was confirmed that the number of syllables in each pseudo-utterance was always five.

The measurement of duration included in the present study is Speech rate in milliseconds (SpRate). Each pseudo-utterance contained five syllables so SpRate was calculated for each pseudo-utterance by dividing total duration by five. In effect, SpRate gives the same significance results as speech rate. Since the number of syllables is constant, the significance results for utterance duration would also have been the same.

7.4 Fundamental frequency (f0)

Whilst English is regarded as having a tendency towards stress-timing (Pike, 1946; Abercrombie, 1967), Japanese has been described as exhibiting a tendency towards isochrony between one mora and another, and therefore labelled a mora-timed language (Ladefoged, 1975; Otake, et al., 1993). Braun and Oba (2007), on the other hand, argue that the term "mora" is still under discussion and that Japanese may in effect be described as a syllable-timed language. Japanese has also been described as a pitch accent language, where pitch accent varies depending upon whether or not a word is stressed and on the number of morae in a word. Pitch accents have been found to distinguish word meaning. There is also regional variation in pitch accent. For further detail, see Cutler and Otake (1999). An attempt was made to control for accent in the encoding experiment. However, there were insufficient participants available from any single area to make this a viable option. There was no control for regional variation in accent in the selection of Japanese or English participants and the data samples were too small to analyse for any regional differences in results.

As explained in Chapter Five, the pseudo-utterances constructed for this study were tested to ensure that they contained no verbal meaning. In addition, each pseudo-utterance contained five single CV syllables with a constant vowel sound for each utterance in order to reduce the possible impact of tendency towards stress-timing in English and mora-timing or syllable-timing in Japanese.

As discussed in Chapter Two, there are various possible scales of measurement of f0. In investigations of the vocal correlates of emotion, f0 is measured in Hertz as

investigators are searching for the phonetic correlates of emotional states. f0 is also generally measured in Hertz in the few cross-cultural studies which have included measurement of f0; for example, in Pell et al. (2009b). With this in mind, since the current study also investigated the phonetic correlates of emotional states, this is what motivated the use of f0 measurements in Hertz here. The author is unaware of any cross-cultural studies of the vocal expression of emotion which investigate the relative merits of different measures of f0 in differentiating the expression of emotions by different cultures. There is scope for future research in this area; however, this is beyond the scope of the current study. Mennen et al. (2008; 2012) is not a study of emotional states. However, they found no difference between the ERB, semitones and Hertz for English in terms of correlations with listener judgements for English and found that ERB and Hertz scales performed better for German data.

In the present study, different measurements were taken in order to try to provide as full a picture as possible of f0 variation with regard to the different affective states. Eight f0 parameters were measured including maximum f0 in Hertz (Maxf0), minimum f0 in Hertz (Minf0), range of f0 in Hertz (Rangef0), Meanf0 in Hertz (Meanf0), standard deviation of f0 in Hertz (SDf0), 90^{th} percentile of f0 in Hertz (90thPercf0), 10^{th} percentile f0 in Hertz (10thPercf0), 80 percent range of f0 in Hertz (80PercRangef0).

As discussed in Chapter Two, methods for measuring f0 do not tend to be detailed in previous studies in the area of emotion in speech. In the current study, whole track f0 values were initially computed by Praat (Boersma & Weenik, 2007). All of the encoders were female so the f0 values were gathered using algorithms implemented in Praat (Boersma & Weenink, 2007), using the settings which are recommended for female voices in the Praat manual.

As in Banziger and Scherer (2005), the f0 track of each pseudo-utterance was manually inspected visually and auditorily by the investigator to check for potential pitch track errors. As discussed in 2.2, there are functions within Praat which will smooth a pitch track or remove octave jumps. However, there is the risk that in attempting to remove anomalous outliers from the f0 tracks by these methods, potentially significant data may be removed so ‘smoothing’ and ‘kill octave jumps’ were not used in the present study. As discussed in 2.2, f0 track errors and how to correct for these is a controversial issue

(Batliner et al., 2007; Steidl et al., 2008) and is a subject of debate between researchers on blogs such as the Praat-users group.

Scheffers (1988) commented that low-pass filtering does not remove from pitch contours irregularities which do not have any relation to the perceived pitch track and that low-pass filtering also affects "the slope and onset and offset moments of the important movements" (1988, p.981). Pfitzinger et al. (2009) claimed that Momel (modelling melody) produces 'a smoothed version that is perceptually indistinguishable from the original and supposedly void of microprosodic fluctuations' (2009, p.2455). Pitch contours were therefore not smoothed in the present study.

The utterances were constructed to have more or less continuous voicing to facilitate pitch contour representation. Utterances tended to be voiced throughout apart from the initial voiceless affricate and sometimes during the hold phase of two voiced plosives. It was therefore not considered necessary to interpolate f0 tracks in order to produce continuous contours as some other studies have done.

No manual correction was therefore made to the f0 track in the present study for the reasons stated above. However, this is an area which requires future investigation. The following list indicates which items would have been potential candidates for adjustment. Comments give more detail for each item.

Few potential candidates for correction were found. These potential errors were at outer values. Inaudible periodicity occasionally occurred in outlying points on the f0 track, but inaudible periodicity also occurred, for example, in the f0 track during the hold phase of voiced stops. Gating and auditory analysis revealed that voicing only became perceptible to the author at the onset of the vowel. This voicing of the hold phase of voiced stops may not have been perceptible to participants in the decoding test but discarding these instances of periodicity was not thought to be justifiable, especially in a cross-cultural study in which decoders, especially with a native language background other than that of the researcher, may have different perceptions of the vocal signal. The explanation for most ‘jumps’ in the data for the present study was found to be creaky voice, which occasionally occurred on the English data for Happy, Sad and Angry. These ‘jumps’ were left uncorrected. There was an instance of periodicity computed in the f0 track of one of the English Angry pseudo-utterances, which was inaudible. Banziger and Scherer (2005) manually unvoiced sections with ‘anomalous’ periodicity. There were four further instances toward the extremes of the pitch track where it was

unclear whether values were errors or not.

It was also sometimes unclear where the f0 track should be shifted to. In addition, not all of these jumps were octave jumps or fifth jumps and would have required manual changing of pathway points. This meant that in whole track measurements at least, creaky voice were sometimes represented in the measurements as pitch level shifts affecting mean and range even though these were not necessarily perceived as creaky voice.

The investigator decided to leave these uncorrected and to measure both whole track and percentile measurements of f0 for all pseudo-utterances. 10^{th} and 90^{th} percentile and 80 percent range measurements were included with the aim of excluding any anomalous outliers which occurred in the upper and lower ten percent of the f0 track. Whole f0 track maximum, minimum, range and mean measurements were also included with the aim of highlighting any differential influence between whole track and percentile measurements upon evidence of distinctions between the vocal cues signalling emotions.

Percentile values were therefore taken to attempt to exclude these errors and whole track and percentile measures were compared to highlight differences in results. It is possible that whole track measurements contained occasional errors. It is also possible that percentile measurements removed significant values. It was anticipated that the percentile measurements may yield more significant distinctions between emotions than the whole pseudo-utterance measurements. However, there was also the possibility that percentile measures may exclude significant data.

The f0 measures used may well influence the findings generated by studies looking at the nature of the vocal correlates of emotion, possibly affecting significant distinctions found between emotions. The use of different f0 measures also has implications when comparing results across studies. This is an area of instrumental phonetics which requires more research generally. The comparative merits of whole track versus percentile measurements for f0 and on whether any correction was made for pitch track errors are not issues which have received attention in investigations of the vocal correlates of emotion. Pell et al. (2009b) is one of the few studies which make reference to the issue of pitch tracking errors. The issue is not explored in depth; however, the authors state they manually corrected for pitch tracking errors and only "…approximately .05% of all tokens" were corrected (Pell et al., 2009b, p.422).

It is often the case in previous studies investigating the vocal correlates of emotion that it is not explicitly stated whether whole track or percentile measurements are taken and if percentile, which percentile measurements were included. Thomson and Balkwill (2006) and Anolli et al. (2008) do state that they calculated f0 mean and f0 range measurements from the whole f0 track whilst Banse and Scherer (1996) for example, state that they used 25th and 75th percentile f0 measurements. However, these studies have not included a comparison between whole track and percentile measurements.

7.4.1 *Maximum, minimum and range of f0*

Two measures of maximum f0 (Hz) and two measures of minimum f0 (Hz) were taken from the 'uncorrected' data. These measures also give two measures for f0 range (Hz). For each pseudo-utterance, whole track maximum f0 and minimum f0 as well as 90th and 10th percentile f0 (Hz) measurements were taken. Taking percentile measurements can help to reduce the possibility of including possibly anomalous outlying frequencies in the data. The percentile measurements were therefore taken for each item with the intention of excluding errors at extremes. However, it was also considered that taking percentile measurements could perhaps result in the exclusion of potentially significant data. The maximum and minimum f0 measurements for the whole pseudo-utterance were included in case omitting the extremes removed data significant to distinctions between emotions and between cultures. The intention was also to highlight similarities and differences between the results derived from taking different types of measurement for f0. Greater consistency between studies in the area of cross-cultural vocal cues of emotion will help if results are to be compared across studies.

7.4.2 *Mean f0 and Standard deviation of f0*

Banse and Scherer (1996) commented that Meanf0 (Hz), as well as being the most prominent perceptually, was the most frequently studied vocal parameter. Meanf0 of each pseudo-utterance (Hz) was measured by accessing 'Periodicity', 'To pitch…' and querying Meanf0 for each pseudo-utterance.

Standard deviation of f0 (Hz) of each pseudo-utterance was measured as an indicator of pitch variability.

7.4.3 *Pitch track start and end points*

How the start points and end points of each pseudo-utterance were decided clearly influenced the values of acoustic measurements and the issues confronted are in need of consideration since more consensus on these issues will result in more accurate comparisons across different studies and will perhaps lead to alternative and more preferable solutions.

Praat screenshots of two of the pseudo-utterances can be found in Appendix E. These screenshots are edited for start and cut off points and show visually where the acoustic measurements were derived from. Figure E1 shows the vocal expression of Fearful on "cha na ga ma ba" by an English encoder and figure E2 is a screenshot of the vocal expression on Japanese Happy on "chee nee gee mee bee" by a Japanese encoder. The starting point of each pseudo-utterance was taken at the start of release of the plosive element of the initial affricate. Finding appropriate and consistent criteria for deciding upon the end cut-off point of the pseudo-utterance proved more difficult.

If the end of final vowel formants were used as an end cut-off point, this sometimes incorporated prolonged breathiness or sighs, which perhaps artificially lengthened the pseudo-utterance. It was decided not to regard these sections as part of the extracted algorithm.

If the algorithm was ended at the end of the pitch contour, auditory checks revealed that the vowel sometimes continued beyond the end of the pitch contour. Using the point at which waveform perturbations returned to pre-pseudo-utterance levels also sometimes missed speech sounds beyond this. Cutting off where low frequency energy ceased (voice striations) clipped some vowel sounds. The method which worked best for the final cut-off point was to ensure that the pitch contour had ended, that the waveform had returned to pre-pseudo-utterance levels, as shown on the screenshots in Appendix E, and an auditory check was also performed. However even this method was not without problems. Some items ended in a post-vocalic glottal stop. In these cases, it was decided to use the start of the close phase of the glottal stop as the end cut-off point. Another item contained additional short pitch contours at the end. An auditory check confirmed that the pseudo-utterance ended perceptually after the first contour so this was where the end cut-off point was made. The formants, contour and waveform of the final vowel in

one item were extended but it was unclear as to where so an auditory check was used to resolve this.

7.5 Intensity

For each pseudo-utterance, whole track and percentile mean and range of intensity measurements were taken. Distinctions between emotions were greater for whole track measurements than for the percentile measurements, which yielded little significance in patterns of distinction between emotions. For brevity, only whole track measurements are included here. These include mean intensity in decibels (MeanInt) and range of intensity in decibels (RangeInt). It should be remembered that the possibility of varying mouth-microphone distance will have influenced measurements of intensity, as discussed in Chapter Five (5.6.2).

7.6 Significant distinctions between emotions for each culture

The data from acoustic analysis was first analysed to test for significant distinctions between emotions for each acoustic parameter in each culture in turn. The data was then analysed to test whether there were significant cross-cultural distinctions in the acoustic characteristics of the English data and the Japanese data. **S**tatistical analysis of acoustic measurements was carried out using the Statistical Package for the Social Sciences (SPSS). In the following discussion comparing significant differences between each emotion for each acoustic parameter in Japanese and English, F ratios and p levels can be found in table form in Appendix F. Table F1 shows distinctions between emotions by acoustic parameter for the English data; Table F2 shows distinctions between emotions by acoustic parameter for the Japanese data.

Whilst not all previous studies have tested for the statistical significance of findings, it was decided that in the interests of facilitating comparison across studies, it would be useful to give a detailed description of tests for significance which were conducted on the data in the present study, including which post-hoc tests were performed to reduce Type I or Type II errors.

For each acoustic parameter means were plotted for each emotion and each culture in order to highlight broad patterns**.** Taking data for each culture in turn, one way

independent analyses of variance (one-way ANOVAs) were calculated using SPSS for each of the acoustic parameters in order to test for any significant variance overall in each acoustic parameter for each emotion for each culture in turn.

The acoustic parameters served as the dependent variables and emotion (5 emotion types) was entered as the independent variable. Results showed normal distribution. The Levene test was performed to test the hypothesis that the variances of each group were equal. If this test showed significance ($p<0.05$), this indicated that the assumption of homogeneity of variance was broken and in this case, the Welch F (1951) test was selected to attempt to resolve the problem. This test controls well for Type I errors, thereby reducing the possibility that a relationship may be classed as significant when it is not. The Welch test is also fairly powerful in reducing Type II errors which occur when relationships are classed as not significant when in fact they are significant. According to Field (2009), the Welch technique is less likely to result in Type II errors than the alternative Brown-Forsythe technique, except where there is an extreme mean which has a large variance. In such instances, Browne-Forsythe was used. Where the Levene test showed that the assumption of homogeneity of variance had been broken, the Welch test results are reported here.

Post-hoc tests were performed for each acoustic parameter in order to access any pair-wise distinctions between emotions. Following the recommendation of Field (2009), since group sizes were unequal, pair-wise comparisons were performed by the use of the Gabriel post-hoc test. This test is more conservative than the LSD post-hoc test which is the equivalent of performing multiple t-tests but reduces the possibility of Type II errors where results may be wrongly classed as not significant. Whilst reducing Type I errors, thereby reducing the possibility that a relationship may be classed as significant when it is not, Gabriel increases the possibility of Type II errors in which relationships are classed as not significant when in fact they are significant. The Gabriel post-hoc test was therefore used to account for Type I errors and the less conservative LSD post-hoc test was used to reduce Type II errors. As would be expected, where significance ($p < 0.05$) was almost reached using the Gabriel post-hoc test, it was always reached in the LSD post-hoc test.

Summaries of significant distinctions between emotions by each acoustic parameter for each culture according to the Gabriel post-hoc test are presented in Tables 7.1 and 7.2 for English and Japanese respectively. These tables are used as a basis for discussion

throughout the rest of this chapter. For clarity, only significant or near significant level results are reported in the tables reporting results of ANOVA tests. Asterisks are used to highlight levels of significance: where p = 0.001 the significance is rated ***. Where p = 0.01 the significance is rated **. Where p = 0.05 the significance is rated *. Where measurements have almost reached significance, the specific level of significance is entered and the significance is rated (*).

Information on the ranked order of emotions for each acoustic variable is given in the final column of Tables 7.1 and 7.2 For example, for Meanf0 within each culture, the emotion with the highest Meanf0 was ranked first and the emotion with the lowest Meanf0 was ranked last. However, it should be noted that only the emotions which are asterisked, without brackets, were found to be significantly distinguished from each other by the indicated acoustic parameter.

Several alternative f0 measurements were taken for mean and range in order to highlight possible anomalies and influence upon distinctions between emotions as discussed in 7.4. Distinctions between English Fearful and other emotions were different depending upon which measures were used for Rangef0. Whole track measurements were shown to yield more distinctions between emotions than percentile measurements. As indicated in Table 7.1, no significant distinctions were found for 10thPercf0. It is possible that the 10thPercf0 measurements excluded some significantly low f0 readings for English Calm, Angry and Sad, leading to the exclusion of significant low f0 values in the 80PercRangef0 readings for these emotions.

For Japanese, Rangef0 showed considerably more distinctions than were shown for 80PercRangef0. This was particularly the case for Japanese Fearful expressions. Rangef0 distinguished Fearful from all of the other emotions in Japanese.

It was noted that distinctions between English Sad and Happy were found for Maxf0 and Rangef0 but not for 90thPercf0 or 80PercRangef0. It is possible that 90thPercf0 or 80PercRangef0 measurements excluded f0 measurements which were significant. On the other hand, where pitch track measurements within the upper 10 percent range were included, these may have indicated results as significant where in fact they were not. As explained in 7.4, none of the potential f0 track errors were corrected either automatically by Praat or manually by the investigator.

Acoustic parameter	Emotion target and significant distinctions from other emotions for the English expressions of emotion using Gabriel post-hoc tests					Order of ranked means
	Happy	**Sad**	**Angry**	**Fearful**	**Calm**	
Sp Rate (sylls per sec)	(*)Sad	(*)Happy				HFCAS Fastest to slowest
Max f0 (Hz)	***Sad; ***Calm	***Happy; **Angry; (*)Fearful	**Sad; **Calm	**Sad; **Calm	*Happy; Angry; **Fearful	HAFSC Highest to lowest
Min f0 (Hz)			(*)Fearful	(*)Angry; *Calm	*Fearful	CASHF Lowest to highest
Range f0 (Hz)	(*)Sad; (*)Fearful	(*)Happy; *Angry;	*Sad; *Fearful; (*)Calm	(*)Happy; *Angry	(*)Angry	AHCSF Widest to narrowest
Mean f0 (Hz)	*Sad; *Calm	*Happy; *Angry; **Fearful	*Sad; *Calm	**Sad; **Calm	*Happy; *Angry; **Fearful	FAHCS Highest to lowest
SD f0 (Hz)						AHCSF Highest to lowest
90th Perc f0 (Hz)	(*)Angry	***Angry; *Fearful	(*)Happy; ***Sad; ***Calm	*Sad; *Calm	***Angry *Fearful	AFHCS Highest to lowest
10thPerc f0 (Hz)						HFSCA Lowest to highest
80 Perc Range f0 (Hz)		*Angry; (*)Fearful	*Sad	(*)Sad		AFHCS Widest to narrowest
Mean Int (dB)	(*)Sad	(*)Happy; *Angry; *Fearful	*Sad	*Sad		AFHCS Highest to lowest
Range Int (dB)						AHFCS Widest to narrowest
SD Int (dB)						AHFCS Highest to lowest

Table 7.1: Significant distinctions between emotions by acoustic parameter for English according to Gabriel post-hoc tests. ***p< 0.001 **p<0.01 *p<0.05 (*)p<0.099. (*) indicates values which are tending towards a significant level. The order of ranked means of each emotion is given for each acoustic parameter. H=Happy; S=Sad; A=Angry; F=Fearful; C=Calm.

Acoustic parameter	Emotion target and significant distinctions from other emotions for the Japanese expressions of emotion using Gabriel post-hoc tests					Order of ranked means
	Happy	**Sad**	**Angry**	**Fearful**	**Calm**	
Sp Rate (sylls per sec)						HCAFS Fastest to slowest
Max f0 (Hz)	(*)Sad; *Calm	(*)Happy; **Fearful		**Sad; **Calm	*Happy; **Fearful	FHASC Highest to lowest
Min f0 (Hz)						FACSH Lowest to highest
Range f0 (Hz)	*Sad; ***Fearful	*Happy; ***Angry; ***Fearful	***Sad; **Fearful; *Calm	***Happy ***Sad; **Angry; ***Calm	*Angry; ***Fearful	FAHCS Widest to narrowest
Mean f0 (Hz)	(*)Sad; *Calm	(*)Happy			*Happy	HFASC Highest to lowest
SD f0 (Hz)		***Fearful		**Calm; ***Sad	**Fearful	FAHCS Highest to lowest
90th Perc f0 (Hz)		*Angry	*Sad; **Calm		**Angry	AFHSC Highest to lowest
10thPerc f0 (Hz)						CSAFH Lowest to highest
80 Perc Range f0 (Hz)	(*)Angry	**Angry	(*)Happy; **Sad; **Calm		**Angry	AFHSC Widest to narrowest
Mean Int (dB)						AFHSC Highest to lowest
Range Int (dB)						SHAFC Widest to narrowest

Table 7.2: Significant distinctions between emotions by acoustic parameter for Japanese according to Gabriel post-hoc tests. ***p< 0.001 **p<0.01 *p<0.05 (*)p<0.099. (*) indicates values which are tending towards a significant level. The order of ranked means of each emotion is given for each acoustic parameter. H=Happy; S=Sad; A=Angry; F=Fearful; C=Calm.

In the English pseudo-utterances, Happy demonstrated the highest Maxf0 whilst Angry had the highest 90thPercf0. It could be that there were f0 extraction errors in the Maxf0 measurements for English Happy. There are differences worthy of note between Minf0 and 10thPercf0 ranking. Whilst Fearful and Happy have the highest Minf0 rankings and Calm and Angry have the lowest Minf0 rankings, this pattern is reversed for 10thPercMinf0. Since Minf0 signals distinctions between Angry, Fearful and Calm and 10thPercMinf0 does not signal any distinctions between emotions, it appears that removing the lower ten percent of f0 track measurements excluded f0 values which underpinned the significant readings from whole track values.

There is then considerable overlap in distinctions between emotions found for whole track f0 and percentile f0 measurements and whole track f0 measures show more distinctions between emotions in the English data than do percentile f0 measurements. For brevity, only whole track f0 measurements are included in the discussion below. However, it should be noted, that 90thPercf0 was the only vocal cue which had any tendency to distinguish Angry from Happy, Angry being expressed with the highest 90thPercf0. Whole track measures of Rangef0 and Maxf0 were also chosen since these facilitated comparison with other studies in this area, which have also tended to be derived from whole utterance measures. Pell et al (2009b) is one of the few studies to correct for pitch tracking errors; however, as discussed above (7.4), they found that only "…approximately .05% of all tokens" (Pell et al., 2009b, p.422) needed to be excluded.

There has been little focus in previous studies in the area of the vocal expression of emotion on the rationale behind which f0 measurements are derived. Both whole track and percentile measurements were included to highlight issues surrounding f0 measurement. The evidence found here of differential results depending upon which measurements of f0 are included highlight the issues surrounding measurement of f0; however, resolution of this issue is beyond the scope of this study.

Tables 7.3 and 7.4 below summarise the distinctions between emotions according to acoustic parameter for English Japanese respectively. All emotions were significantly distinguished from each other in both cultures except for a lack of distinction between the vocal cues of Sad and those for Calm. This was possibly due to both of these emotions falling within the category of low arousal. Despite Calm and Happy falling within the same valence category, both having a positive valence, Calm was found to

have a distinctively lower Maxf0 and Meanf0 than Happy in the English expressions and a lower Maxf0 and Meanf0 than Happy in the Japanese expressions.

High arousal emotions were distinguished from low arousal emotions by many of the acoustic parameters by both cultures. Whilst most distinctions were between high and low arousal emotions, significant distinctions were also found between high arousal emotions.

Sections 7.7 and 7.8 present in more detail the distinguishing characteristics found for each emotion for English and for Japanese respectively. Results for each culture are then compared (7.9). F ratios and p levels can be found in table form in Appendix F. Table F1 shows distinctions between emotions by acoustic parameter for English; Table F2 shows distinctions between emotions by acoustic parameter for Japanese data and Table F3 shows the main effect for culture and for culture plus emotion.

7.7 Vocal characteristics of the English expressions

This section explains the pattern of vocal characteristics found for each emotion in the English pseudo-utterances and whether these vocal cues distinguished each emotion from the other emotions investigated.

Table 7.3 below illustrates the vocal cues of each emotion for English. A value for level is given where the specific acoustic parameter distinguishes an emotion from at least one other. Opposing levels between emotions for a particular acoustic parameter do not necessarily indicate that these levels are significantly distinct. For significant distinctions between emotions see the descriptions for each emotion above and in Appendix F, Table F1.

Descriptions in brackets indicate results tending towards significance. Mid is given as a description only if an emotion is distinguished (or is tending towards distinction) from one emotion by having a higher/wider level and from another emotion by having a lower/narrower level.

Acoustic parameter	Acoustic correlates of emotions in English pseudo-utterances				
	Happy	**Sad**	**Angry**	**Fearful**	**Calm**
SpRate (sylls per sec)	(Fast)	(Slow)			
Maxf0 (Hz)	High	Low	High	(High)	Low
Minf0 (Hz)			(Low)	High	Low
Rangef0 (Hz)	(Wide)	Narrow	Wide	Narrow	(Narrow)
Meanf0 (Hz)	High	Low	High	High	Low
SDf0 (Hz)					
MeanInt (dB)	High	Low	High	High	
RangeInt (dB)					

Table 7.3: Significant acoustic correlates of emotions in English pseudo-utterances according to Gabriel post-hoc tests. Descriptions in brackets indicate results tending towards significance. Where no description is given, no distinction was found. Mid is given as a description only if an emotion is distinguished (or is tending towards distinction) from one emotion by having a higher/wider level and from another emotion by having a lower/narrower level.

Vocal cues were found in the English expressions of emotion which distinguished emotions from each other except for a lack of distinction between Calm and Sad. This was possibly due to both of these emotions falling within the category of low arousal. Despite Calm and Happy falling within the same valence category, both having a positive valence, Calm was found to have a distinctively lower Maxf0 and Meanf0 than Happy for English.

As can be seen from Table 7.3, high and low arousal emotions were distinguished by many of the acoustic parameters in the English data. Whilst most distinctions were between high and low arousal emotions, significant distinctions were also found between high arousal emotions. English Happy was expressed with a distinctively wider Rangef0 than English Fearful. Japanese Angry was expressed with a distinctively wider Rangef0 than English Fearful. English Fearful was distinguished from the other high arousal emotions by having a significantly narrower Rangef0 than Angry, also tending towards a significantly narrower Rangef0 than Happy and tending towards a significantly higher Minf0 than Angry.

Post-hoc tests for each acoustic parameter revealed the following statistically significant distinctions between emotions for English.

7.7.1 *English Happy*

SpRate, f0 and intensity were all relevant to the distinction of English Happy from the other emotions. English Happy pseudo-utterances were characterised by a fast SpRate, high Maxf0, wide Rangef0, high Meanf0 and high MeanInt, all of which were significant or tended towards significance in relation to at least one of the other emotions. Taken together, these vocal cues distinguished Happy expressions from Sad and Calm in particular and tended towards distinguishing Happy from Angry and Fearful. There is some redundancy suggested since several cues signal distinction between Happy and Sad and Happy and Calm.

English Happy was distinguished from Sad and Calm particularly by its high Maxf0 but also by a significantly higher Meanf0 and MeanInt. Happy was also expressed with a SpRate which tended towards being significantly faster than Sad and a Rangef0 which tended towards being significantly wider than Rangef0 for Sad or Fearful. Happy was expressed with a relatively higher MeanInt than Sad, which tended towards significant distinction.

7.7.2 *English Sad*

SpRate, f0 and intensity were all relevant to the distinction of English Sad from Happy, Angry and Fearful. English Calm was not distinguished from English Sad by the parameters investigated.

English Sad was expressed with a slow Speech rate, low Maxf0, narrow Rangef0, low Meanf0, low MeanInt and a narrow RangeInt and together these vocal cues distinguished English Sad expressions from Happy, Angry and Fearful expressions. There was also redundancy of vocal cues distinguishing Sad from the high arousal emotions. However, none of the acoustic parameters investigated showed significant distinction between Sad and Calm expressions. Similarities between the vocal cues of Calm and Sad would be expected since these are both low arousal emotions.

English Sad was expressed with a slow Speech rate, tending towards distinction from Happy; a low Maxf0 which strongly distinguished Sad from Happy and Angry expressions and tended towards distinguishing Sad from Fearful expressions; a significantly narrower Rangef0 than English Angry and tending towards being

significantly narrower than English Happy; a significantly lower Meanf0 than Fearful, Angry and Happy and a low MeanInt which distinguished Sad from Angry and Fearful and tended towards distinction from Happy.

7.7.3 *English Angry*

f0 and intensity were relevant to the distinction of English Angry from the other emotions. English Angry was expressed with a high Maxf0, low Minf0, wide Rangef0, high Meanf0 and a high MeanInt, all of which were significant or tending towards significance in distinguishing Angry from one or more of the other emotions. Taken together, these vocal cues distinguished Angry expressions from Sad and Calm in particular. Angry was also distinguished from Fearful and tended towards distinction from Happy. There is some redundancy suggested, particularly in cues signalling distinction between Angry and Sad and Angry and Calm.

English Angry was distinguished from Sad and Calm particularly by its high Maxf0, but also by a significantly higher Meanf0 for Angry than for Sad or Calm. A high MeanInt tended towards distinguishing Angry from Sad expressions.

A wide Rangef0 distinguished Angry not only from Sad but also from Fearful and there was a tendency towards distinction from Calm by Rangef0. Apart from distinction between Angry and Fearful by Rangef0, Angry having the widest Rangef0 and Fearful the narrowest, Angry also tended towards distinction from Fearful by the related parameter of Minf0, Angry expressions having a much lower Minf0 than Fearful expressions.

7.7.4 *English Fearful*

f0 and intensity were relevant to the distinction of English Fearful from the other emotions. English Fearful was expressed with a high Maxf0, high Minf0, narrow Rangef0, high Meanf0 and high MeanInt, all of which were significant or tended towards significance and taken together these vocal cues distinguished Fearful expressions from Sad and Calm in particular. Fearful was also distinguished from Angry and tended towards distinction from Happy. There is some redundancy suggested, particularly in cues signalling distinction between Fearful and Sad and

Fearful and Calm.

English Fearful was distinguished from Sad and Calm particularly by its high Maxf0 and high Meanf0. Fearful was also distinguished from Sad by Fearful expressions having a distinctively higher MeanInt. A high Minf0 distinguished Fearful from Calm and tended to distinguish Fearful from Angry and a narrow Rangef0 distinguished Fearful from Angry and tended to distinguish Fearful from Happy.

7.7.5 *English Calm*

f0 characteristics were found to be relevant to the distinction of English Calm from Happy, Angry and Fearful. English Calm was not distinguished from English Sad by any of the parameters investigated. English Calm was expressed with a low Maxf0, low Minf0, narrow Rangef0 and a low Meanf0 all of which were significant or tended towards significance and together these vocal cues distinguished English Calm expressions from Happy, Angry and Fearful expressions. In common with English Sad, there was also redundancy of vocal cues distinguishing Calm from the high arousal emotions. However, none of the acoustic parameters investigated showed significant distinction between Calm and Sad expressions. Similarities between the vocal cues of Calm and Sad would be expected since these are both low arousal emotions.

English Calm was expressed with a low Maxf0 which distinguished Calm expressions from Happy, Angry and Fearful expressions; a low 90thPercf0 level which distinguished Calm from Angry and from Fearful; a narrow Rangef0 which tended towards significant distinction from Angry expressions and a significantly lower Meanf0 than Fearful, Angry and Happy. Whilst measures of intensity distinguish Sad from Angry and Fearful in the English expressions, intensity cues did not distinguish Calm from any of the other emotions.

7.8 Vocal characteristics of the Japanese expressions of emotion

This section explains the pattern of vocal cues found for each emotion in the Japanese pseudo-utterances and whether these vocal cues distinguished each emotion from the other emotions investigated.

Vocal cues were found in the Japanese expressions of emotion which distinguished emotions from each other except for a lack of distinction between Calm and Sad. This was possibly due to both of these emotions falling within the category of low arousal. Despite Calm and Happy falling within the same valence category, both having a positive valence, Calm was found to have a distinctively lower Maxf0 and Meanf0 than Happy for Japanese.

Acoustic parameter	**Acoustic correlates of emotions in Japanese pseudo-utterances**				
	Happy	**Sad**	**Angry**	**Fearful**	**Calm**
SpRate (sylls per sec)					
Maxf0 (Hz)	High	Low	High	High	Low
Minf0 (Hz)					
Rangef0 (Hz)	Mid	Narrow	Mid	Wide	Narrow
Meanf0 (Hz)	High	(Low)			Low
SDf0 (Hz)		Low		High	Low
MeanInt (dB)					
RangeInt (dB)					

Table 7.4: Significant acoustic correlates of emotions in Japanese pseudo-utterances according to Gabriel post-hoc tests. Descriptions in brackets indicate results tending towards significance. Where no description is given, no distinction was found. Mid is given as a description only if an emotion is distinguished (or is tending towards distinction) from one emotion by having a higher/wider level and from another emotion by having a lower/narrower level.

As can be seen from Table 7.4, high and low arousal emotions were distinguished by many of the acoustic parameters in Japanese. Whilst most distinctions were between high and low arousal emotions, significant distinctions were also found between high arousal emotions. Japanese Happy was expressed with a distinctively narrower Rangef0 than Japanese Fearful. Japanese Angry was expressed with a distinctively narrower Rangef0 than Japanese Fearful. Japanese Fearful demonstrated a wide Rangef0 which distinguished this emotion from all other emotions.

Table 7.4 illustrates the vocal cues of each emotion for Japanese. A value for level is given where the specific acoustic parameter distinguishes an emotion from at least one other. Opposing levels between emotions for a particular acoustic parameter do not necessarily indicate that these levels are significantly distinct.

Descriptions in brackets indicate results tending towards significance. Mid is given as a description only if an emotion is distinguished (or is tending towards distinction) from one emotion by having a higher/wider level and from another emotion by having a

lower/narrower level.

Vocal cues were found in the Japanese expressions of emotion which distinguished emotions from each other except for a lack of distinction between Calm and Sad. This was possibly due to both of these emotions falling within the category of low arousal. Despite Calm and Happy falling within the same valence category, both having a positive valence, Calm was found to have a distinctively lower Maxf0 and Meanf0 than Happy for Japanese.

7.8.1 *Japanese Happy*

Only f0 parameters were found to show significant distinctions between Japanese Happy and the other emotions.

Japanese Happy was expressed with a faster Speech rate than any other emotion, although the rate was not found to be significantly faster that of the other emotions. Happy was expressed with a high Maxf0, mid Rangef0, and a high Meanf0, all of which were significant or tended towards significance and taken together these vocal cues distinguished Happy expressions from Sad and Calm in particular. There is some redundancy suggested since several cues signal distinction between Happy and Sad and Happy and Calm. Happy was also found to be distinguished from Fearful and tended towards distinction from Angry.

Japanese Happy was expressed with higher Maxf0 and a higher Meanf0 than Sad and Calm expressions. These distinctions were significant or tended towards significance. Happy expressions also had a significantly wider Rangef0 than Sad. Happy was also expressed with a Rangef0 which was highly significantly wider than for Fearful. Neither of the intensity cues was significant in distinguishing Happy from the other emotions. Happy was expressed with a mid MeanInt and a high RangeInt.

7.8.2 *Japanese Sad*

Only f0 parameters were found to show significant distinctions between Japanese Sad and the high arousal emotions. No significant distinctions were found between Sad and Calm.

Japanese Sad was expressed with a slower Speech rate than any other emotion, although the rate was not found to be significantly slower that of the other emotions. Japanese Sad was expressed with a low Maxf0, narrow Rangef0, low Meanf0, low level of variability in f0 (SDf0), low MeanInt and the widest RangeInt. These distinctions were significant or tended towards significance and together these vocal cues distinguished Japanese Sad expressions from Happy, Angry and Fearful expressions. There was also redundancy of vocal cues distinguishing Sad from the high arousal emotions. None of the acoustic parameters investigated showed significant distinction between Sad and Calm expressions. However, note the comments below under Japanese Calm.

7.8.3 *Japanese Angry*

Only f0 parameters were found to show significant distinctions between Japanese Angry and the other emotions.

Japanese Angry was expressed with a high Maxf0 and mid Rangef0, both of which were significant or tended towards significance in distinguishing Angry from one or more of the other emotions. Taken together these vocal cues distinguished Angry expressions from Sad and Calm in particular. There is some redundancy suggested in cues signalling distinction between Angry and Sad and Angry and Calm. Angry was distinguished from Fearful by Rangef0.

7.8.4 *Japanese Fearful*

Only f0 parameters were found to show significant distinctions between Japanese Fearful and the other emotions.

Japanese Fearful was expressed with the highest Maxf0, the widest Rangef0, and the highest level of SDf0, which were all significant. Taken together these vocal cues distinguished Fearful expressions from Sad and Calm in particular. There is some redundancy suggested, particularly in cues signalling distinction between Fearful and Sad and Fearful and Calm. The wide f0Range of Japanese Fearful distinguished the expression of Japanese Fearful from the expression of each of the other emotions.

7.8.5 *Japanese Calm*

Only f0 parameters were found to show significant distinctions between Japanese Calm and the high arousal emotions. No significant distinctions were found between Calm and Sad.

Japanese Calm was expressed with the lowest Maxf0, a narrow Rangef0, the lowest Meanf0, and a low SDf0, all of which were significant and together these vocal cues significantly distinguished English Calm expressions from Happy, Angry and Fearful expressions. In common with Japanese Sad, there was redundancy of vocal cues distinguishing Calm from the high arousal emotions. None of the acoustic parameters investigated showed significant distinction between Calm and Sad expressions.

Section 7.9 includes discussion of main effect for culture and for culture plus emotion, according to each acoustic parameter.

7.9 Cross-cultural comparison of emotion distinctions by acoustic parameter

Additional one-way ANOVAs testing for the significance of culture were conducted for each emotion in turn with each acoustic parameter as the dependent variable and culture as the fixed factor in order to highlight overall levels of distinction between the English and Japanese in the levels of each acoustic parameter. The Levene test was used to test homogeneity of variance and where this test showed significance (i.e. non-homogenous variation was found), the more robust Welch F values were entered. The data was tested for main effect for culture and main effect for culture and emotion. Gabriel post-hoc tests were performed to test for significant distinctions between the levels of acoustic parameters for English and Japanese and to test for significant distinctions between English and Japanese in measurements of each acoustic parameter for each emotion. In the present study, only the association between SpRate and emotion varied significantly as a function of language. Mean SpRate was significantly faster for the Japanese data than for the English data overall, $F(1,40) = 7.131$, $p = 0.011$.

Further distinctions emerged between the English and Japanese data when the mean levels for each acoustic parameter were analysed in relation to each emotion in turn. In the following discussion comparing significant differences between mean levels of each acoustic parameter for English and Japanese for each emotion, F ratios and p levels

showing main effect for culture and emotion plus culture can be found in table form in Appendix F, Table F3.

Whilst for English, measurements of fundamental frequency and intensity were significant in distinguishing between emotions and there was a tendency towards distinction by speech rate, for the Japanese data, emotions were distinguished by only measurements of fundamental frequency.

For English, overall, the measurements of fundamental frequency found to be significant included Maxf0, $F(4,16) = 11.762$, ***p = 0.000, Meanf0, $F(4, 16) = 9.242$, ***p = 0.000, Rangef0, $F(4, 16) = 5.782$, **p = 0.004 and Minf0, $F(4, 16) = 3.839$, *p = 0.023. MeanInt was also found to be significant in distinguishing between emotions for English, $F(4, 16) = 3.813$, *p = 0.023. Emotions expressed by the English encoders tended towards distinction by SpRate, $F(4,16) = 2.461$, (*)p = 0.87.

In the Japanese data, distinctions between emotions were signalled by Maxf0, $F(4,6.698) = 11.762$, **p = 0.0049, Rangef0 (Hz), $F(4, 7.609) = 202.176$, ***p = 0.000 and SDf0 (Hz), $F(4, 6.481) = 20.049$, ***p = 0.001. Emotions tended towards distinction by Meanf0 (Hz), $F(4, 6.718) = 4.015$, (*)p = 0.056.

Research into the vocal cues of emotion has tended to find the strongest evidence for distinctions between the vocal cues used for high and low arousal emotions See Table 3.1). Taking high and low arousal emotion as a base, English and Japanese cues are compared for each acoustic parameter and each emotion. Given the tendency towards lower levels of identification according to target emotion for the Japanese vocal expressions, it was predicted that the Japanese vocal cues would demonstrate less salient characteristics than English cues with regard to expected cues for high and low arousal emotions. Expected cues for high and low arousal emotions are based on evidence in previous studies, as discussed in Chapter Two. So, for example, mean f0 has been found to be higher for high arousal than for low arousal emotions so it was predicted that mean f0 would be higher for English Angry, and therefore more salient, than for Japanese Angry and lower for English Sad, and therefore more salient, than for Japanese Sad.

The following sections therefore include discussion of distinctions between emotions by each acoustic parameter for English and for Japanese. There is discussion of any distinctions between English and Japanese overall for each acoustic parameter. There is

also explanation of any interaction between each emotion and each acoustic parameter for each culture and a cross-cultural comparison of this evidence. The findings are illustrated visually by the graphs. Estimated marginal means (EMMeans) are illustrated rather than simple means in order to take account of unequal group sizes. There is also some discussion of how the findings could help to explain decoding test results.

7.9.1 *Speech Rate (syllables per second)*

ANOVA tests revealed that SpRate was significantly slower in English than in Japanese for Calm ($F(1,8) = 6.409$, $^*p = 0.035$) and was slower for in English than in Japanese for Sad at a level approaching significance ($F(1,8) = 5.214$, $(^*)p = 0.052$)

The association between SpRate and emotion overall was the only parameter to vary significantly as a function of culture. Pellegrino, Coupé and Marsico, (2011) also report evidence of a faster speech rate for Japanese. It is possible that culturally distinct 'vocal settings' for SpRate could be a significant factor in explaining cross-cultural differences in the vocal cues of emotion. Braun and Oba (2007) also reported that Japanese vocal expressions of emotion tended to be faster than American English and German expressions.

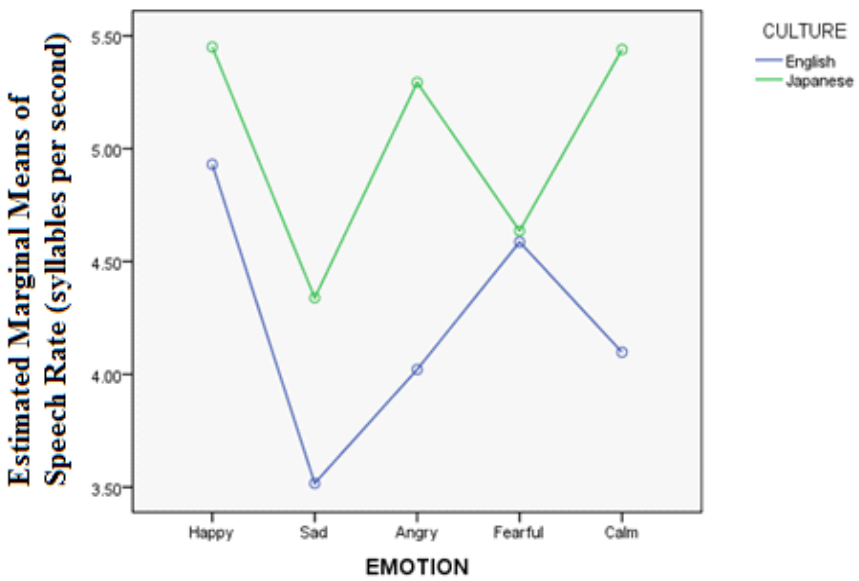

Figure 7.1: A comparison between SpRate (syllables per second) used by English and Japanese encoders in the expression of Happy, Sad, Angry, Fearful and Calm. (Estimated Marginal Means were derived to take account of unequal group sizes.)

As can be seen in Figure 7.1, each of the English emotions was expressed with a slower SpRate than the Japanese expressions, although there was little difference for Fearful. There was a tendency towards distinction between Happy and Sad by SpRate in English, Happy being expressed with a faster SpRate. Whilst descriptively Happy had the fastest and Sad the slowest SpRate in Japanese, the difference was not revealed as significant.

There is evidence in previous studies of a distinctively faster SpRate for Happy than for Sad. The distinction approaches significance for English in this study and no significant distinctions were found between any of the emotions for SpRate in Japanese. The faster SpRate for Sad in Japanese could have contributed to the lower recognition levels for Japanese Sad compared to English Sad. The utterances used in the present study were only five syllables long. The utterances used in the present study were only five syllables long. This is not unusually short for studies in this area which investigate SpRate. As discussed, in Erikson et al. (2008a), for example, the three-syllable word 'banana' was used as a stimulus. Nevertheless, significant differences in SpRate were found for culture and emotion in both the study by Erikson et al. (2008a) and in the present study as discussed.

7.9.2 *Maximum f0 (Hz)*

As discussed in sections 7.6 and 7.7, high arousal emotions (Happy, Angry and Fearful) were expressed with a significantly higher Maxf0 than low arousal emotions (Sad and Calm) in the English acoustic data. For Japanese, high arousal emotions (Happy and Fearful) were expressed with a significantly higher Maxf0 than low arousal emotions (Sad and Calm).

Figure 7.2 shows that Maxf0 was higher for Happy and Angry in English than in Japanese. However, this parameter did not significantly distinguish Angry from the four other emotions in Japanese. Angry was also expressed with a significantly higher Maxf0 in English than in Japanese ($F(1,5) = 10.222$, $*p = 0.024$). Since high Maxf0 is a cue which has often been found related to Angry, this evidence could help to explain the lower decoding levels for Angry in the Japanese data.

Maxf0 was found to be significantly higher in Japanese than in English for Sad (F(1,8)

= 8.137, *p = 0.021). Since there is considerable evidence in previous studies of low Max f0 as a cue for Sad, the lower Maxf0 in the English data for Sad would provide a more salient cue of Sad in English than in Japanese, again perhaps helping to explain the lower decoding levels for the Japanese expressions. The higher decoding levels of the English data than of the Japanese data for Sad even by the Japanese decoders suggests that the lower Maxf0 for Sad could be a psycho-physiological cue of which Japanese decoders are aware, even if this is not produced by the encoders in Japanese.

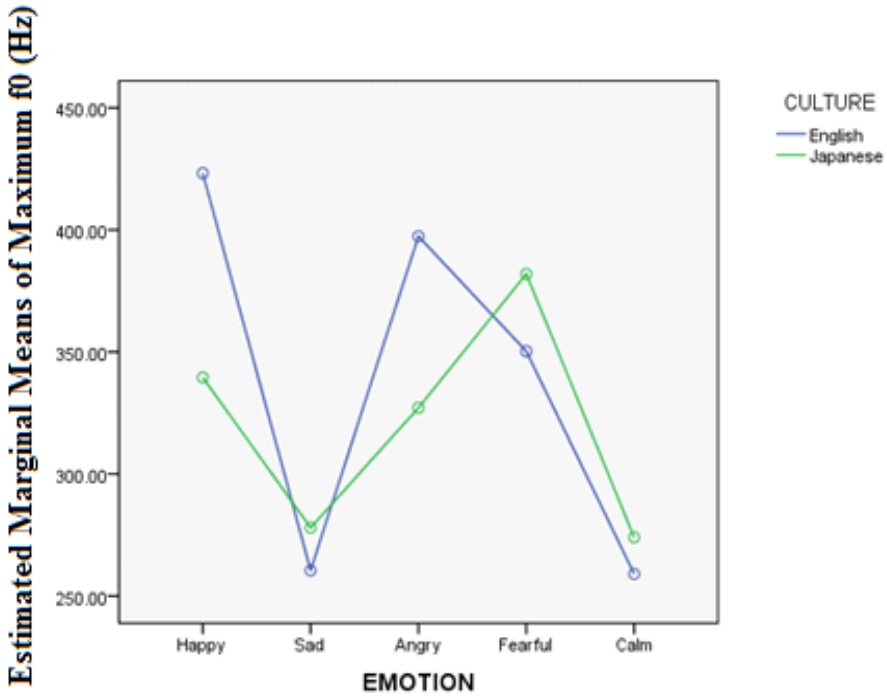

Figure 7.2: A comparison between Maximum f0 (Hz) used by English and Japanese encoders in the expression of Happy, Sad, Angry, Fearful and Calm. (Estimated Marginal Means were derived to take account of unequal group sizes.)

Evidence of High Maxf0 has also been found in previous studies to be a cue for Happy. The tendency towards significance in the higher Maxf0 for Happy expressions in English compared to Japanese could suggest a possible tendency towards greater salience in this cue for Happy in English than in Japanese.

The distinguishing cue of Maxf0 was possibly more perceptibly salient for Happy, Angry and Sad in the English pseudo-utterances than in the Japanese pseudo-utterances, which could help the lower decoding levels found for the Japanese data for these emotions.

These results suggest that the Japanese vocal cues for Happy and Angry were less salient than the English cues if we predict high Maxf0 for high arousal emotions.

7.9.3 *Minimum f0 (Hz)*

Fearful was expressed with by far the highest Minf0 in English and had a significantly higher Minf0 than Anger ($F(4, 16) = 3.839$, $p = 0.082$) and Calm ($F(4, 16) = 3.839$, $p = 0.03$).

Figure 7.3 shows that Fearful was expressed with the lowest Minf0 in Japanese and Happy and Sad were expressed with the highest Minf0. However, no significant distinctions between emotions were indicated by this acoustic parameter for Japanese.

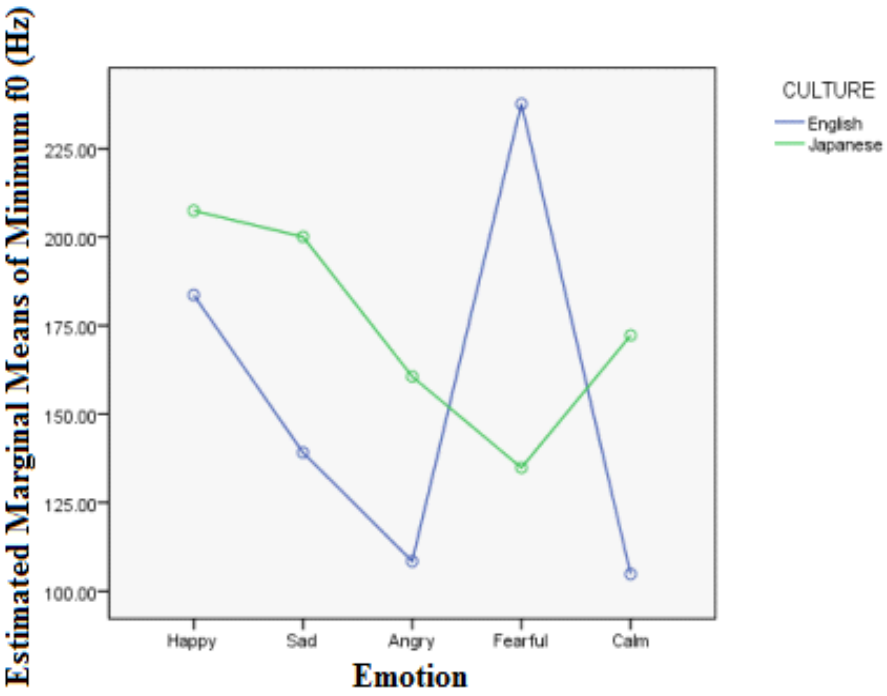

Figure 7.3: A comparison between Minimum f0 (Hz) used by English and Japanese encoders in the expression of Happy, Sad, Angry, Fearful and Calm. (Estimated Marginal Means were derived to take account of unequal group sizes.)

Visually, the means plot in Figure 7.3 shows mean Minf0 is higher in English than in Japanese for all emotions except for Fearful which is very much higher in English than in Japanese. ANOVAs showed distinction approaching significance for Fearful ((*)$p = 0.058$). There was a tendency towards significance in the higher Minf0 for Fearful

expressions in English compared to Japanese. No other emotions showed significant distinction between English and Japanese.

7.9.4 *Range f0 (Hz)*

English Fearful was distinguished from the other high arousal emotions by having a significantly narrower Rangef0 than Angry ($F(4, 16) = 5.782$, $p = 0.016$), also tending towards a significantly narrower Rangef0 than Happy and tending towards a significantly higher Minf0 than Angry as shown in Table F1 of Appendix F.

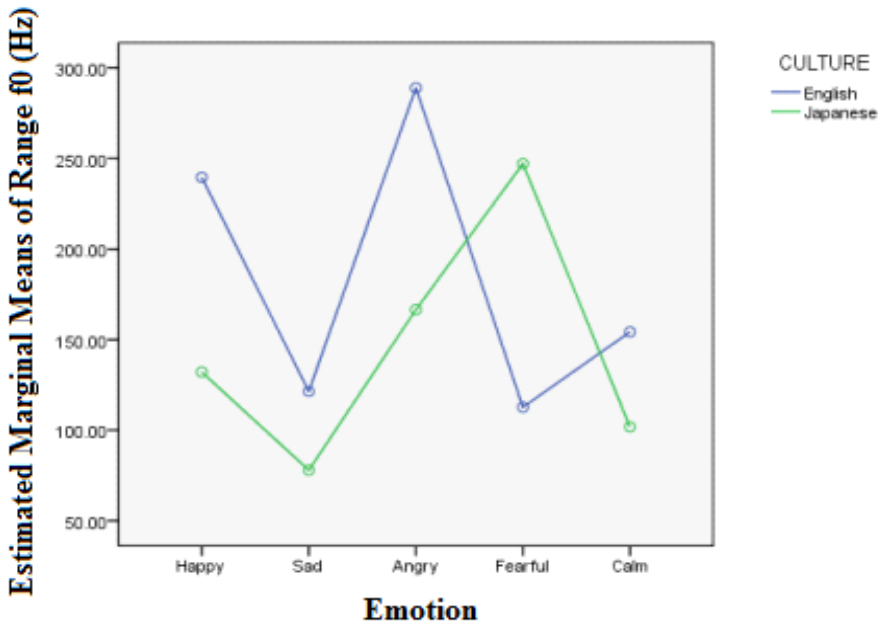

Figure 7.4: A comparison between Range f0 (Hz) used by English and Japanese encoders in the expression of Happy, Sad, Angry, Fearful and Calm. (Estimated Marginal Means were derived to take account of unequal group sizes.)

For English, as shown in Table F1 of Appendix F, Rangef0 for Anger was significantly wider than that for Fearful, Sad or Calm. Rangef0 also distinguished Happy from Sad and Fearful. Angry was expressed with a significantly wider Rangef0 in English than in Japanese, as can be seen in Appendix F, Table F2.

English Happy was expressed with a distinctively wider Rangef0 than English Fearful, whilst in contrast Japanese Happy was expressed with a distinctively narrower Rangef0

than Japanese Fearful.

Japanese Angry was expressed with a distinctively narrower Rangef0 than Japanese Fearful and with a distinctively wider Rangef0 than English Fearful. Japanese Fearful demonstrated a wide Rangef0 which distinguished this emotion from all other emotions.

There was a tendency towards significance in the wider Rangef0 for Happy and Calm expressions in English compared to Japanese and Fearful was expressed with a significantly wider Rangef0 in Japanese than in English (Table F3). Happy was expressed with a significantly wider Rangef0 than Sad in Japanese (Table F2).

Figure 7.4 illustrates that Fearful was expressed with by far the widest Rangef0 in Japanese and Happy, Calm and particularly Sad, were expressed with the narrowest range. Table F2 shows that the Rangef0 for Fearful was significantly wider than that for Angry, Happy, Calm and Sad and Rangef0 also distinguished Happy from Sad and Fearful. Angry was also expressed with a significantly wider Rangef0 than Sad and Calm.

The means plot in Figure 7.4 shows that the English data had a wider Rangef0 for all emotions except for Fearful, which had a significantly wider range in Japanese (Table F3). As Table F3 illustrates, Angry was expressed with a significantly wider Rangef0 in English than in Japanese, which provides evidence that this cue could have been a more salient cue in English than Japanese since previous studies have found evidence of wide Rangef0 as a cue for Angry. There was no significant difference found between English and Japanese in levels of Rangef0 used to express Sad.

7.9.5 *Mean f0 (Hz)*

ANOVAs showed a distinction between English and Japanese in Meanf0 for Sad which was approaching significance. Sad was expressed with a significantly lower Meanf0 in English than in Japanese (Table F3). There is previous evidence of low Meanf0 for Sad, so again, this could have helped to cause the English vocal expression of Sad to be more reliably decoded. However, no main effect for culture and emotion was found for Meanf0 for any of the other emotions. This is the inverse of results found for Rangef0.

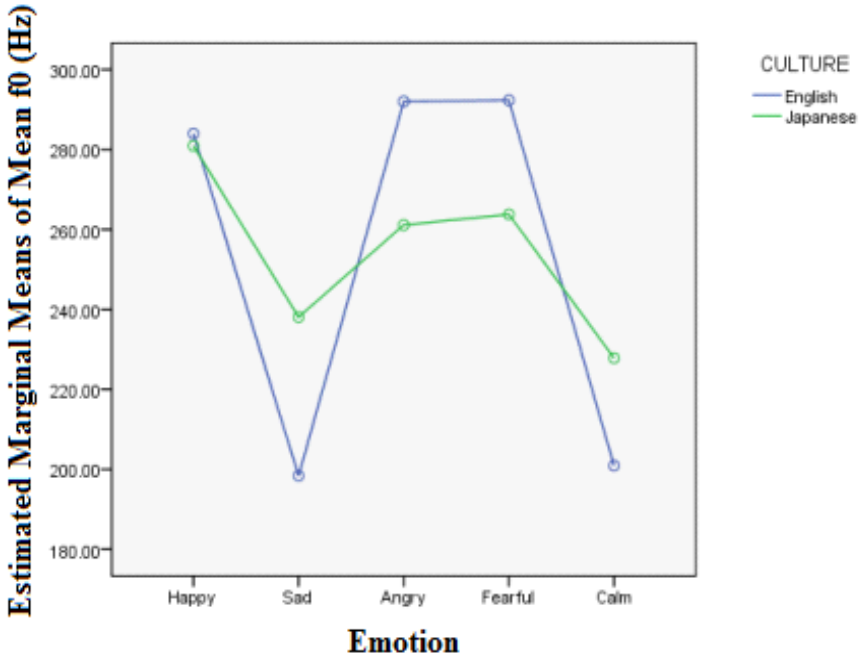

Figure 7.5: A comparison between Mean f0 (Hz) used by English and Japanese encoders in the expression of Happy, Sad, Angry, Fearful and Calm. (Estimated Marginal Means were derived to take account of unequal group sizes.)

In the English expressions, high arousal emotions (Happy, Angry and Fearful) were expressed with a similar Meanf0 levels. As shown in Table F1, were significantly higher than low arousal emotions (Sad and Calm), which had a similar Meanf0 levels. For Japanese, Happy was expressed with the highest Meanf0 and the Meanf0 level for Happy was significantly higher than that for Calm and Sad (Table F2). Meanf0 did not distinguish Angry or Fearful in the Japanese data. There is considerable previous evidence of a high Meanf0 as a cue for high arousal emotions. The evidence in this study suggests that whilst the English uses high Meanf0 to cue Angry and Fearful, there is no evidence of this in the Japanese data. Again, this may help to explain the higher decoding levels for English Angry than for Japanese Angry, even by Japanese decoders. It is argued that this provides tentative evidence of high Meanf0 being a psych-physiological cue for Angry.

7.9.6 *Standard deviation f0 (Hz)*

Figure 7.6: A comparison between Standard Deviation f0 (Hz) used by English and Japanese encoders in the expression of Happy, Sad, Angry, Fearful and Calm. (Estimated Marginal Means were derived to take account of unequal group sizes.)

Whilst visual observation of Figure 7.6 shows Happy and Angry having a much higher level of standard deviation than Calm, Sad or Fearful for English, no significant distinctions were found in the data for this acoustic measure for the English data. However, there was a tendency towards significance in the higher SDf0 for Angry expressions in English compared to Japanese. If high SDf0 is a cue for Angry, this would make the English data more salient for this emotion by this acoustic parameter.

Visual observation of the graph in Figure 7.6 shows that Japanese Fearful is expressed with a much higher level of SDf0 than the other four emotions and statistical analysis showed Fearful to have significantly greater standard deviation than Calm and Sad (Table F2). Fearful was expressed with a significantly higher SDf0 in English than in Japanese (Table F3). It is possible that high SDf0 is a vocal cue for Fearful which is recognised cross-culturally as the English decoders identified Japanese Fearful according to target at least as well as they identified English Fearful.

7.9.7 *Intensity Range (dB)*

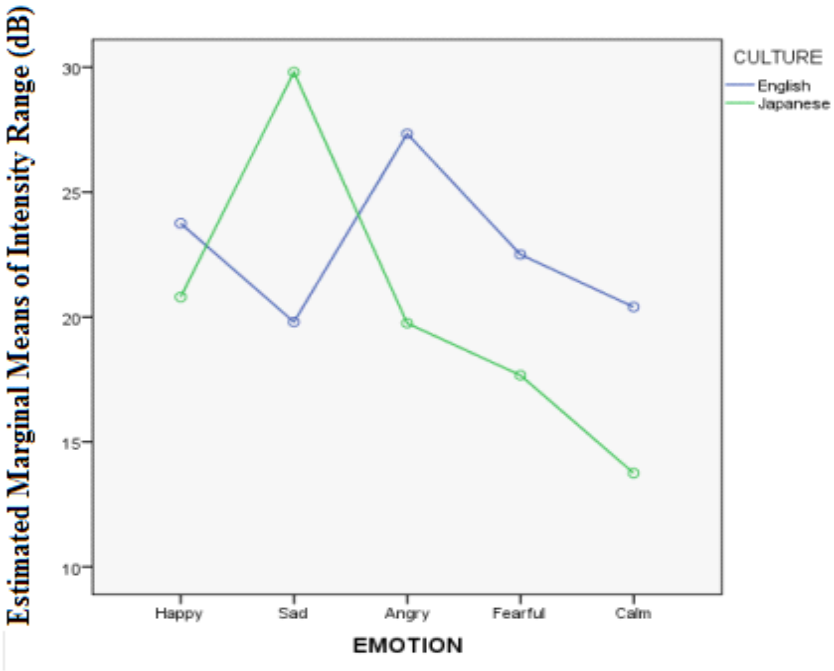

Figure 7.7: A comparison between Intensity Range (dB) used by English and Japanese encoders in the expression of Happy, Sad, Angry, Fearful and Calm. (Estimated Marginal Means were derived to take account of unequal group sizes.)

No significant differences were found for the association between RangeInt and each emotion for the English data or for the Japanese data. However, ANOVAs showed a distinction between English and Japanese in Intensity Range (dB) for Calm which was approaching significance. Calm was vocalised with a narrower RangeInt in Japanese than in English. This could suggest that the Japanese data was more salient for Calm by this acoustic parameter. However, lack of previous evidence for Calm means that this is a very tentative observation. In addition, Japanese Sad, which is the other low arousal emotion in the set, was vocalised with a wider RangeInt than English Sad or any of the other emotions vocalised by the Japanese encoders.

Since there is evidence in previous studies that a wider RangeInt tends to express high arousal emotions, a narrower Rangef0 being used for low arousal emotions, the pattern found here would suggest that the English data is more salient for Happy, Sad, Angry and Fearful.

7.9.8 *Mean Intensity (dB)*

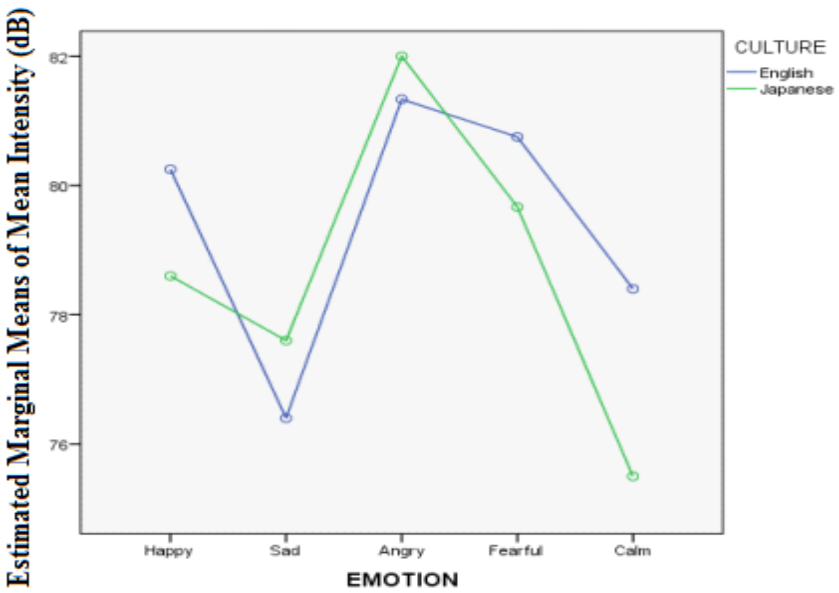

Figure 7.8: A comparison between Mean Intensity (Hz) used by English and Japanese encoders in the expression of Happy, Sad, Angry, Fearful and Calm. (Estimated Marginal Means were derived to take account of unequal group sizes.)

No interaction was found between culture and MeanInt for emotion overall. As expected, the high arousal emotions Happy, Angry and Fearful were expressed with higher MeanInt in both cultures and Angry was expressed with the highest MeanInt than other emotions in both English and Japanese.

7.10 Summary

For each culture, vocal cues significantly distinguished emotions except for a lack of distinction between the vocal cues of Sad and Calm. This was possibly due to both of these emotions falling within the category of low arousal. Despite Calm and Happy falling within the same valence category, both having a positive valence, Happy expressions were expressed with a significantly higher Maxf0 and Meanf0 than Calm expressions by both cultures (Tables F1 and F2).

High arousal emotions were distinguished from low arousal emotions by many of the

acoustic parameters by both cultures. Whilst most distinctions were between high and low arousal emotions, as discussed above, significant distinctions were also found within the set of high arousal emotions.

Very similar patterns were found for English and for Japanese in the vocal cues used to signal the emotions investigated. However, there were more significant distinctions between the vocal cues which signalled the emotions in English than in Japanese and there is evidence that vocal cues are more salient for English than for Japanese as explained in 7.9 above. These findings therefore provide support for the claim of pan-cultural vocal cues of emotion. In addition, evidence was found in the present study that the level of salience and distinctiveness of these cues is culture-dependent. It is possible that pitch accent could have had an effect on the f0 characteristics of the Japanese vocal expressions. It is also possible that the Japanese encoders were more inhibited by the experiment than the English encoders, which could also have led to less salient cues.

As can be seen from Table F3, only the association between SpRate and emotion varied significantly as a function of language. However, further distinctions emerged between the English and Japanese data when the mean levels for each acoustic parameter were analysed in relation to each emotion in turn. There was a tendency towards significance in the wider Rangef0 for Happy expressions in English compared to Japanese. Sad was expressed with a significantly lower Maxf0 and Meanf0 in English than in Japanese. There was a tendency towards significance in the slower SpRate for Sad expressions in English compared to Japanese. Angry was expressed with a significantly higher Maxf0 and wider Rangef0 in English than in Japanese. There was a tendency towards significance in the higher SDf0 for Angry expressions in English compared to Japanese. Fearful was expressed with a highly significantly wider Rangef0 in Japanese than in English and was expressed with a significantly higher SDf0 in Japanese than in English. There was a tendency towards significance in the lower Minf0 for Fearful expressions in English compared to Japanese. Calm was expressed with a significantly slower SpRate in English than in Japanese and there was a tendency towards significance in the wider Rangef0 for Calm expressions in English compared to Japanese.

These distinctions indicate that the English data often tended to be more salient than the Japanese data in terms of what would be expected in relation to high and low arousal patterns found in this study and previous studies, although the evidence for Fearful and Calm is less clear. There is conflicting evidence as to whether narrow or wide Rangef0

cues Fearful and it appears that there could be cross-cultural variation.

We are some way down the track from having a sense of what the relevant acoustic cues of even basic emotions are within given speech communities. It should be recognised that variations in emotion vocalisation and perception will also tend to occur within the same culture as well as cross-culturally. However, to date, studies have tended to search for common patterns, without focussing on individual variation in either decoding results or in acoustic analysis. It is still very early days for cross-cultural research in this area and this remains an area for future investigation within the field of the vocal expression of emotion. In the small number of studies which have performed acoustic analysis on data encoding vocally expressed emotion, there tends to be a small number of subjects, sometimes only one. Even in studies with more encoding subjects from the same culture, individual variation in the vocal expression of emotion has not tended to be analysed. The present study is the first to use a balanced, symmetrical methodology in the still new area of cross-cultural vocalisation of emotion. It is an early small-scale study which has taken a broad-brush approach rather than focussing on individual differences, which is beyond the scope of this study. In the next chapter, the findings of acoustic analysis are discussed in relation to evidence of in-group and cross-cultural decoding and conclusions are drawn.

Chapter Eight

Discussion and Conclusions

This study was set up with four research questions and a balanced, symmetrical methodology was constructed to address these, including also acoustic analysis of the vocal expressions by encoders from each culture. The symmetrical methodology of this study is a novel element and represented a significant challenge. However, this methodology provides a strong basis on which to answer the research questions of the study. It is in line with the framework of Scherer's adaptation of The Brunswikian Lens Model (Scherer, 2003), in that the study encompasses consideration of both production of 'distal' vocal cues, analysed by acoustic analysis, and perception of 'proximal' vocal cues by the inclusion of in-group and cross-cultural decoding tests.

As an initial step, the **first research question** investigated whether for each culture, emotions could be recognised from vocal cues, and whether or not some emotions were more reliably identified according to target than others. In-group decoding of all emotions was found to be better than chance for both cultures, supporting previous research. For both cultures, Happy, Sad and Angry were found to be relatively well-recognised in-group compared to Fearful and Calm. As can be seen in Tables F1 and F2 (Appendix F), fundamental frequency, particularly Rangef0, was found to be significant for both cultures in distinguishing between the emotions investigated, except for the two low arousal emotions, Sad and Calm. Calm was, however, distinguished from Happy, the other positive valence emotion in the set. Happy demonstrated a significantly higher Maxf0 and Meanf0 than Calm for both cultures. However, Japanese Happy was most often confused with Angry and Calm. This could be predicted by Happy and Angry both being high arousal emotions and Happy and Calm both having a positive valence.

It was also found that Japanese decoders showed lower levels than English decoders for in-group decoding. This supports previous reports that East Asian cultures tend to show more restraint than Western cultures in the expression of emotion. There is a need for further research across diverse cultures to confirm whether or not this is generally the case.

The **second research question** examined the extent to which vocal cues could signal emotion cross-culturally and beyond chance recognition demonstrated that Japanese and English participants were able to identify Happy, Sad, Angry, Fearful and Calm from

vocal cues alone, although Japanese identification of English Fearful was borderline with chance level (24%). English Fearful was very often confused with Sad, both in-group and cross-culturally and the narrow Rangef0 for English Fearful could have led to the confusion with English Sad which was also characterised by a narrow f0. In contrast, Japanese Fearful was expressed with a wide Rangef0. Previous research has found conflicting results for the vocal cues of Fearful in particular and this is often attributed to different types of Fearful being vocalised (e.g. anxiety or panic). There is evidence in this study that different vocal cues are used for Fearful cross-culturally, although it is also possible that different types of Fearful were being produced, despite the symmetry of the encoding experiment. Further cross-cultural research will be needed to ascertain the extent of cross-cultural variation in the expression of Fearful and to tease out whether different cues are due to different types of Fearful being produced or are based on cross-cultural variation.

Mainly negative emotions have been included in previous studies, due to most emotions classed as basic emotions being negative. The inclusion of more than one positive valence emotion allowed investigation of characteristics of under-researched positive valence emotions. Confusions tended to follow dimensional lines of arousal, although there is some evidence of confusion between emotions of common valence. High arousal Happy, Angry and Fearful tended to be confused with each other rather than with low arousal Calm or Sad and vice versa. An exception was English Fearful, which was often heard in-group as Sad; Japanese decoders were twice as likely to hear English Fearful as Sad. This is possibly influenced by Fearful and Sad both being negative valence emotions.

Calm was identified according to target beyond chance by both cultures, both in-group and cross-culturally and was generally confused most with the only other low arousal emotion, Sad. In fact, Japanese decoders were more likely to identify both Japanese and English Calm as Sad. This supports previous evidence of confusion between emotions of similar arousal, suggesting primacy given to the arousal over the valence dimension; vocal correlates of Calm were found to be similar to those of Sad for both cultures and none of the acoustic parameters investigated showed significant distinction between Calm and Sad expressions in either the English or the Japanese data and further research with additional parameters is needed here. Whilst Happy was more reliably identified than Calm, English decoding of Japanese, Happy was less well recognised than Calm. There is also tentative evidence of Japanese decoders confusing English Calm with

Happy. These findings could have been due to both these emotions having a positive valence. Further research is needed into positive emotions in studies which, like the present study, include more than one positive emotion.

Significant distinctions between English and Japanese in the mean level of each acoustic variable for each emotion and overall can be found in Table F3 of Appendix F. No overall interaction between culture and emotion as a whole was found for the fundamental frequency and intensity parameters in the present study. However, interaction for specific emotions and culture were found for f0 parameters. For example, Angry was expressed with a significantly wider Rangef0 in English than in Japanese and Fearful was expressed with a distinctively wider Rangef0 in Japanese than English. One parameter which distinguished Japanese Fearful from Sad was its wide Rangef0. The distinctively narrower Rangef0 of English Fearful could at least partly explain why Japanese decoders were around twice as likely to identify the English Fearful target as Sad.

One question for future investigation is whether culturally distinct vocal settings may be responsible for these differences. Whilst the present study did not set out to enter this new area by testing for culturally differential influence of vocal settings on the expression and recognition of emotion, the observation that speech rate was found to be slower overall for the English data than for the Japanese data in this study of emotional speech, raised the possibility that speech rate may be slower in general in English than in Japanese. Previous studies have not tested for the potential influence of culturally differential vocal settings on the vocal cues of emotion.

Evidence was found for cultural differentiation in the speech rate for Sad and for Calm. Sad tokens identified according to target were expressed with a slower SpRate in English than in Japanese and the difference approached significance. If, as has often been found in research into the vocal expression of emotion, Sad is expressed with a SpRate which is distinctively slower than that of high arousal emotions, it is possible that a faster vocal setting for SpRate in Japanese would make the vocal cues of Sad less distinctive and therefore less recognisable. In this study, SpRate approached significance in distinguishing English Happy and Sad, Happy having a faster SpRate. However, SpRate did not distinguish any emotions in the Japanese encoded data.

It is then possible that particular vocal settings may make some emotions easier to recognise and some more difficult to decode. For example, the generally slower speech

rate in English compared to Japanese may help to explain why both English and Japanese decoders recognised the English expression of Sad more reliably than they recognised the Japanese expression of Sad. However, it should be noted that there was little difference in overall decoding rates for English and Japanese Calm, which was the other low arousal emotion in the set. Similarly, a faster overall speech rate in Japanese did not appear to improve recognition of high arousal emotions (Happy, Angry and Fearful) in Japanese compared to English.

Whilst there appears to be cultural influence on vocal settings, it is not possible to deduce from the evidence in the present study, the extent to which covariational influences may play a role. For example, it is possible that there is a certain range of speech rate which co-varies with quasi-universal psycho-physiological emotion responses and this is an area for future research.

The possible influence of vocal settings on expression and recognition of emotion is an area which merits investigation in the field of cross-cultural production and perception of vocal cues of emotion. Previous research, which has not been concerned with investigating the vocal cues of emotion, such as Keating and Kuo (2010), Mennen et al. (2012) and Yamazawa and Hollien (1992), compared cross-cultural vocal settings for f0 and found evidence for cross-cultural differences. There is also evidence that Japanese has a generally faster speech rate (syllables per second) than English (Pellegrino, Coupé & Marsico, 2011). The small-scale study by Braun and Oba (2007) reported that Japanese vocal expressions of Joy, Sad, Angry and Fearful tended to be faster than American English and German expressions in their study.

There is then to date a lack of research focussed on the possible influence of cross-cultural differences in 'vocal settings' on the vocal cues of emotion and in the recognition of these cues. It is possible that culturally distinct 'vocal settings' for speech rate, pitch, intensity and potentially voice quality could also be significant factors in helping to explain cross-cultural differences in the vocal cues of emotion, and this could be a fruitful area for future studies.

The **third research question** explored evidence of In-Group Advantage in the two cultures investigated. Evidence has been found in this study of In-Group Advantage in the recognition of vocal cues of emotion. However, In-Group Advantage is both emotion-dependent and culture-dependent, In-Group Advantage having been found for English decoding of Happy, Fearful and Calm and for Japanese decoding of Happy. It

has been suggested (Tooby & Cosmides, 1990) that In-Group Advantage in the recognition of Happy has an evolutionary explanation, since, it is argued, Happy is important for social cohesion and is less relevant to relations outside the social group.

It is perhaps possible that In-Group Advantage may be off-set by cultural cues of emotion being acquired or consciously learnt by members of different cultures, however close or distant their cultures are considered to be. This possibility does not appear to have been considered in previous studies. This phenomenon could be termed 'Cultural Cue Awareness' (CCA) in order to highlight its potential influence. In terms of Scherer's adaptation of The Brunswikian Lens Model (Scherer, 2003), whilst we might expect that children learn to adapt their covariant psycho-physiological responses to cultural display rules as they grow and integrate within a particular culture, it is also possible that humans learn to differentially conjure these responses depending upon their interaction with other cultures. It is also possible that the degree to which individuals adapt these responses to different cultures may vary. There may be a multitude of reasons for such variation. Further research may help to ascertain possible patterns.

There is also evidence in the present study which appears to directly contradict the concept of In-Group Advantage, in that cross-cultural recognition was more reliable than in-group recognition for particular emotions. Japanese decoders recognised the English vocal cues of Happy and Sad more reliably according to target than they recognised the vocal cues of these emotions by Japanese encoders.

English Happy was distinguished from Sad by a significantly higher Maxf0, Meanf0, and MeanInt. In addition, the wider Rangef0 and faster SpRate approached levels of significant distinction. On the other hand, Japanese Happy was distinguished from Sad by fewer parameters, and all of the distinctions related to f0. The Japanese cues also tended to be less salient, as discussed in Chapter Seven. It appears that whilst Japanese vocal expressions did not exhibit the same vocal cues as English they were able to recognise these cues. It is argued that this may be due to recognition of vocal cues which are influenced by psycho-physiological mechanisms, which can be recognised pan-culturally even if they are not produced by a particular culture. This new concept has been coined here as 'Psycho-Physiological Cue Awareness' (PPCA).

Previous studies have occasionally found evidence of more reliable recognition of emotion cross-culturally than in-group and this was attributed to differences in decoding

conditions between one culture and another. The balanced methodology and symmetrical encoding procedures of the present study mean that the evidence cannot be explained in this way. The more reliable recognition of English Happy and Sad than Japanese Happy and Sad provides evidence of an effect which appears to go directly against the effect of In-Group Advantage.

This suggests that the Japanese decoders were capable of recognising vocal cues of emotion which they do not use themselves, providing tentative evidence for the existence of acoustic cues of Happy and Sad in the English expressions which may be quasi-universally recognised even if, possibly due the influence of cultural display rules, they are not necessarily universally expressed, perhaps due to different levels of restraint related to emotional expression in the display rules of different cultures. Even though a decoder may restrain the expression of these cues themselves due to cultural display rules, they may well still be aware of these cues. This effect may be termed 'Psycho-Physiological Cue Awareness' (PPCA). In theory, the concept of PPCA may apply not only to the voice but also to other communication modes which may be psycho-physiologically influenced such as facial expression and body gesture.

The concept of PPCA may be related to Scherer's adaptation of The Brunswikian Lens Model (2003) applied to the vocal communication of emotion, which was discussed in Chapter One and has been referred to in previous chapters. This model is useful in helping to explain PPCA, since it includes the full process of emotion generation and communication, encompassing universal psycho-physiological ('covariational') cues which are triggered by appraisal of an event. The model suggests that these cues may be adapted and adjusted due to cultural display rules. It is only one step further to suggest that a listener would be aware of these cues, aligned as they are to psycho-physiological processes they also deal with, even if they are more likely to regulate these processes in order to control the outward expression of their emotion through vocal cues.

If indeed Japanese decoders recognised acoustic cues of basic emotions in the English data which were the result of quasi universal psycho-physiological influence on the autonomic nervous systems, this would suggest that covariational influence was less inhibited by cultural display rules in English. Equally, it is possible that in English there are fewer configurational effects than in Japanese.

An alternative or possibly additional explanation for Japanese decoders' high decoding accuracy of English emotion targets could be that the Japanese participants may have

been more aware of the vocal correlates of emotion used by native speakers of English than originally anticipated, possibly due to exposure through the global mass media.

Another possible explanation for English expressions being more reliably decoded both in-group and cross-culturally than Japanese expressions is that the English and Japanese participants may have been influenced differently by the experimental procedure itself. For example, Japanese encoders may have been more reticent than English encoders to produce the emotions in a laboratory setting. The Japanese encoders were also given instructions for the experiment by someone from a different culture and in a language with which they had some familiarity, but which was not their own native language.

Conversely, it is possible that cultural cues may be particularly prevalent for a specific culture, as could be evidenced by high in-group decoding levels and low decoding levels found consistently for cross-cultural decoders of these cues from other cultural groups. The concept of CCA is relevant here to encompass the idea that these cues could potentially be acquired or learnt, however close or distant the culture of the decoder is from that of the encoder. Again this concept may be applied to cues in other non-verbal modes of communication, such as facial expression and gestural cues.

The cues resulting from greater or lesser salience of psycho-physiological cues are also in a sense, cultural cues since salience could vary from culture to culture. It does not appear to be the case that the Japanese vocal expressions in the present study include strong cultural cues to the emotions expressed vocally since they are decoded with around the same level of accuracy by both Japanese and English decoders. It could be that emotions in Japanese are expressed more through verbal, facial or gestural modes than through the vocal channel.

As discussed in Chapter One, it has been suggested (Scherer, 1986; Scherer, 2013a) that the covariation model may be more applicable to basic emotions and that the configurational model is more adequate for describing the processes behind cognitive affect states such as ‘doubt’. Evidence in the present study suggests that the extent of covariational and configurational influence in the expression of basic emotions is culturally dependent. As explained in Chapters One and Three, the Appraisal Model (Ellsworth & Scherer, 2003; Scherer, 1984; 1986; Scherer & Ellgring, 2007) focuses on individual evaluation of an event by stimulus evaluation checks according to various criteria. Individual variation, including the potential role of the individual conscious mind, is an important area still to be explored.

However, evidence of cross-cultural similarity in the vocal cues of basic emotions in English and Japanese supports the theory that there are links between appraisal of an event, the physiological substratum and the vocal response (Johnstone et al, 2001; Scherer, 1986; Scherer, 2013a), also supporting Scherer's adaptation of The Brunswikian Lens Model (Scherer, 2003).

It is possible that PPCA reflects 'push effects' or covariational effects and In-Group Advantage reflects 'pull effects' or configurational effects. These effects were discussed in Chapter One. Evidence of both PPCA and In-Group Advantage has been found in this study. It is argued that the concept of PPCA could provide an explanation of results in future studies, which may otherwise be difficult to account for, or may be explained as anomalies. It illustrates the usefulness of a balanced design and symmetrical experimental procedures in revealing potential influences and effects which could otherwise remain undiscovered. A balanced methodology in cross-cultural studies, including symmetrical encoding procedures, is suggested as a useful vehicle for furthering understanding of the cultural and pan-cultural influences upon the vocal correlates of emotion.

The **fourth research question** investigated the possible influence of vowel quality on in-group and cross-cultural decoding of emotion. There is some support for the theory that [u] can signal aggression (Ohala, 1984; Xu & Chuenwattanapranithi, 2007) as Angry was more likely to be decoded on [u]. In-group decoding of English Fearful was also much lower on [u] (19%) than on [i] (69%) or [a] (60%). Evidence was found that Happy may be easier to identify on the spread vowel [i], which could suggest a smile (van Bezooijen, 1984; Xu & Chuenwattanapranithi, 2007). In addition, where English decoders confused English Calm expressed on [i] with another emotion, they tended to decode Calm as Happy, whilst they confused Calm much more with Sad on the other two vowel qualities. However, support was not found in in-group Japanese decoding of Happy, which was more well recognised on [a] than on [i] or [ɯ]. Results for Japanese Happy and for the Japanese close back, unrounded vowel [ɯ] pose questions which merit further investigation. English decoding of Japanese Happy on [i] was at 63%. However, whilst [ɯ], like [i], is a spread vowel and may also be suggestive of a smile, English decoders' were four times more likely to confuse Happy on [ɯ] with the other positive emotion, Calm. Perhaps the degree of lip-spreading or the front characteristic of [i] was influential, the less well-identified Japanese vowel [ɯ] being a back vowel. Based on these results, further cross-cultural research is merited on vowel quality.

The present study has investigated the role of 11 acoustic parameters relating to fundamental frequency (Hz), intensity (dB) and speech rate (syllables per second) in the vocal expression of emotion by speakers from one Western and one Eastern culture. This is an unusually large number of parameters to include in a cross-cultural study. Voice quality and fundamental frequency contour, for example, are potentially also relevant and have started to be considered, but remain an area in need of more research, particularly in cross-cultural studies. However, they are beyond the scope of this investigation. It is argued that the inclusion of auditory analysis in future studies in this area could provide useful insights into the proximal vocal cues of emotion. This would fit within the framework of Scherer's adaptation of The Brunswikian Lens Model (Scherer, 2003), in relation to proximal cues (2.2).

The few previous cross-cultural studies in this area which have included acoustic analysis have tended to search for evidence of cross-cultural similarities and differences in vocal cues of emotion for which evidence has been found in mono-cultural studies. These parameters include various measurements for fundamental frequency (Hz), intensity (dB) and speech rate (syllables per second), which relate to the prosodic features of pitch, loudness and timing. The scope of the present study was also limited to these cues. Nonetheless, a total of 11 acoustic parameters were analysed, which relate to fundamental frequency (Hz), intensity (dB) and speech rate (syllables per second). Investigation of the influence of other acoustic parameters such as voice quality and fundamental frequency contour upon the vocal correlates of emotion is sparser in cross-cultural research. Further investigation of the role of these parameters in the expression of emotion in different cultures could provide further evidence of distinctive vocal profiles of specific emotions.

Given the need for further research into the communication of emotion as an indexical function of speech, and in particular the extent to which this is psycho-physiologically or culturally influenced, this study has focussed on cross-cultural comparison of ability to decode emotion from in-group and cross-cultural vocal cues and analysis of what these cues may be. The results suggest both psycho-physiological and cultural influence on the vocal cues of emotion. Firm conclusions cannot be drawn regarding the influence of culture and pan-cultural psycho-physiology upon the recognition of emotion from vocal expressions, nor regarding the acoustic cues involved, until more cross-cultural research of a wide variety of cultures is undertaken. Future research would benefit from further balanced, symmetrical studies to test the findings presented here in relation to

vowel influence and PPCA and the inclusion of auditory analysis could provide more understanding of proximal cues. However, the new concept of PPCA and cross-cultural evidence of vowel influence based on evidence found in this thesis have opened up new avenues of investigation in the area of cross-cultural vocal cues of emotion. It has also been suggested that fruitful results could be gained from focussed investigation of the influence of culturally differential vocal settings on cross-cultural recognition of vocal cues of emotion.

The symmetrical, balanced methodology used in this study could be replicated to investigate vocal cues of emotion used in other cultures, which would facilitate more direct comparison between studies investigating different cultures.

Multimodal studies will be needed to investigate the interplay between different modalities in the communication of emotion. Cross-cultural comparison of the interplay between different modalities will be required to form a fuller picture of the communication of emotion, although this will provide substantial methodological challenges for future research.

The findings which have led to the formation of the concept of 'Psycho-Physiological Cue Awareness' (PPCA) demonstrate an ability to decode certain vocal signals of emotion even if these signals are not used by a particular culture. Future research will be needed to indicate whether this phenomenon is generalised across more cultures and whether there is evidence of PPCA in other modes. This study has also coined the term 'Cultural Cue Awareness' (CCA) to refer to awareness of culture-specific cues. This phenomenon is related to In-Group Advantage and the Cultural Proximity Hypothesis. However, CCA is different in that this term encompasses the idea that cultural cues of emotion can be acquired or learnt by members of different cultures, however close or distant their cultures are considered to be. Since the development of greater CCA has a vital role to play in improving cross-cultural communication, an additional perspective in future studies investigating the cues of emotion in vocal, verbal, facial and gestural modes could be to focus on the extent to which culture-specific cues are learnt by members of different cultures. This could potentially contribute to increasing consciousness of differences in cultural cues of emotion, whilst highlighting the possibility that people from different cultures can acquire or learn each other's cues.

Given the importance of emotion to human experience, there are potentially many applications of findings of emotion research, such as cross-cultural communication,

child language acquisition, speech and language communication disorders and computer-based systems for human-computer interaction and clinical psychology. Cowie (2011) has called for a critical eye to be kept on the ethical implications of the application of findings, commenting that "Fulfilling moral obligations tends not to be easy, but ignoring them should not be an option" (p.711).

Whilst considerably more cross-cultural research will be needed, including further balanced, symmetrical studies and investigation of individual variation, before stronger conclusions can be drawn, this study has provided evidence of both cultural influence and cross-cultural psycho-physiological influence on the vocal cues of basic emotions. From a broader, interdisciplinary perspective, further insights will be gained by considering both these influences, but also the potential role of the individual conscious and unconscious mind and body, and by viewing all these aspects, not as separate, but as vitally connected, interacting influences in the experience, understanding and communication of emotion.

Appendices

Note

Unless otherwise stated, materials in Appendices A, B and C were those presented to native English-speaking participants. Where the word ‘English’ is used the word ‘Japanese’ was substituted in materials presented to native Japanese-speaking participants. Where the emotion labels Happy, Sad, Angry, Fearful and Calm occur in materials presented to native English-speaking participants, the labels ureshi-i, kanashi-i, atamanikuru, kowai, and ochitsuku respectively are substituted in the materials presented to native Japanese-speaking participants.

Appendix A: Participant profile questionnaire

1. First name(s) _____
2. Last name(s) _____
3. Contact telephone number _____

   E-mail _____
4. Sex: male female
5. Age _____
6. Which countries have you lived in?

7. In which region(s) of which country have you spent most of your life?

8. Present occupation:

   Past occupation(s):
9. What is your first/native language?
10. Do you speak any other languages? If so, to what level?

    language level (e.g. beginner, intermediate, advanced)

11. Please indicate below if you have had significant contact with people who speak a different native/first language:

    language length of time (approximate) situation

12. Have you watched many foreign films? (E.g. American films if you are not American.) If so, please indicate language(s):

Please use the other side of this paper if you need more space.
Thank-you for your help.

Appendix B: Pre-test

Name_____

Cover **Sections 2 and 3**. Complete **Section 1.**
Cover **Section 3**. Complete **Section 2**.
Complete **Section 3.**
Complete **Section 4.**

Section 1
Look at each photo and decide which emotion is being expressed.

Use more than one word if you wish. Write down any sounds the English person in the photo or you might make if they/you were feeling this emotion.

Photo	**Emotion**	**sound(s)**
1		
2		
3		
4		
A		
B		
C		
D		

Section 2
If each numbered photo represents the same emotion as one of the lettered photos, which numbers go with which letters? Which emotion do they both express?

number	**letter**	**emotion**
1		
2		
3		
4		

Please let me know when you have finished these two questions. Do not alter your answers to these questions whilst you are answering question 3.

Section 3

You will be given 4 words. If you had to attach one of these to each number/letter pair, which word would you attach to which number/letter? Which sounds would you attach to each number/letter pair?

number	**letter**	**word**	**sound(s)**
1			
2			
3			
4			

Section 4

Which sounds might you make or expect another English person to make if you/they were feeling "calm"?

What makes you feel calm? e.g. a particular activity, a piece of music, a situation, an image, something else…?

If possible, describe an experience you remember which made you feel calm.

Thank-you for your help

Appendix C: Reliability test

Name_____

You will hear 80 examples of people speaking with different emotions.

a. Please indicate which of the following emotions you think they are expressing:

happiness

sadness

anger

fear

calm

b. You will then be told which emotion the speaker was trying to express.

Please indicate how convincing you think their expression was on a scale of 1 - 5:

5 - excellent

4 - good

3 - satisfactory

2 - almost satisfactory

1 - unsatisfactory

C. Make any other comments you wish to make.

1.

2.

(Participant test sheet numbered 1-80)

Thank-you for your help.

Appendix D: Decoding test

Name_____

In order for this experiment to be valid, each participant must complete this questionnaire on their own.

You will hear examples of people speaking with different emotions.

Only 5 different emotions are expressed:

happy sad angry fearful calm

For each example:

1. Circle **which emotion** you think the person is expressing. For example:

happy sad angry fearful calm

PLEASE CIRCLE ONLY ONE EMOTION FOR EACH EXAMPLE.

You will not understand the language spoken. Your decisions must be based on the sound of the voices you hear.

2. Indicate **how sure** you are that you have made the right choice. For example:

not sure sure very sure

PLEASE CIRCLE ONLY ONE RESPONSE FOR EACH EXAMPLE.

Practice examples

1. happy sad angry fearful calm not sure sure very sure

2. happy sad angry fearful calm not sure sure very sure

(Participant test sheet numbered 1-90)

Thank-you for your help

Appendix E: Screenshots of Praat Analysis Windows

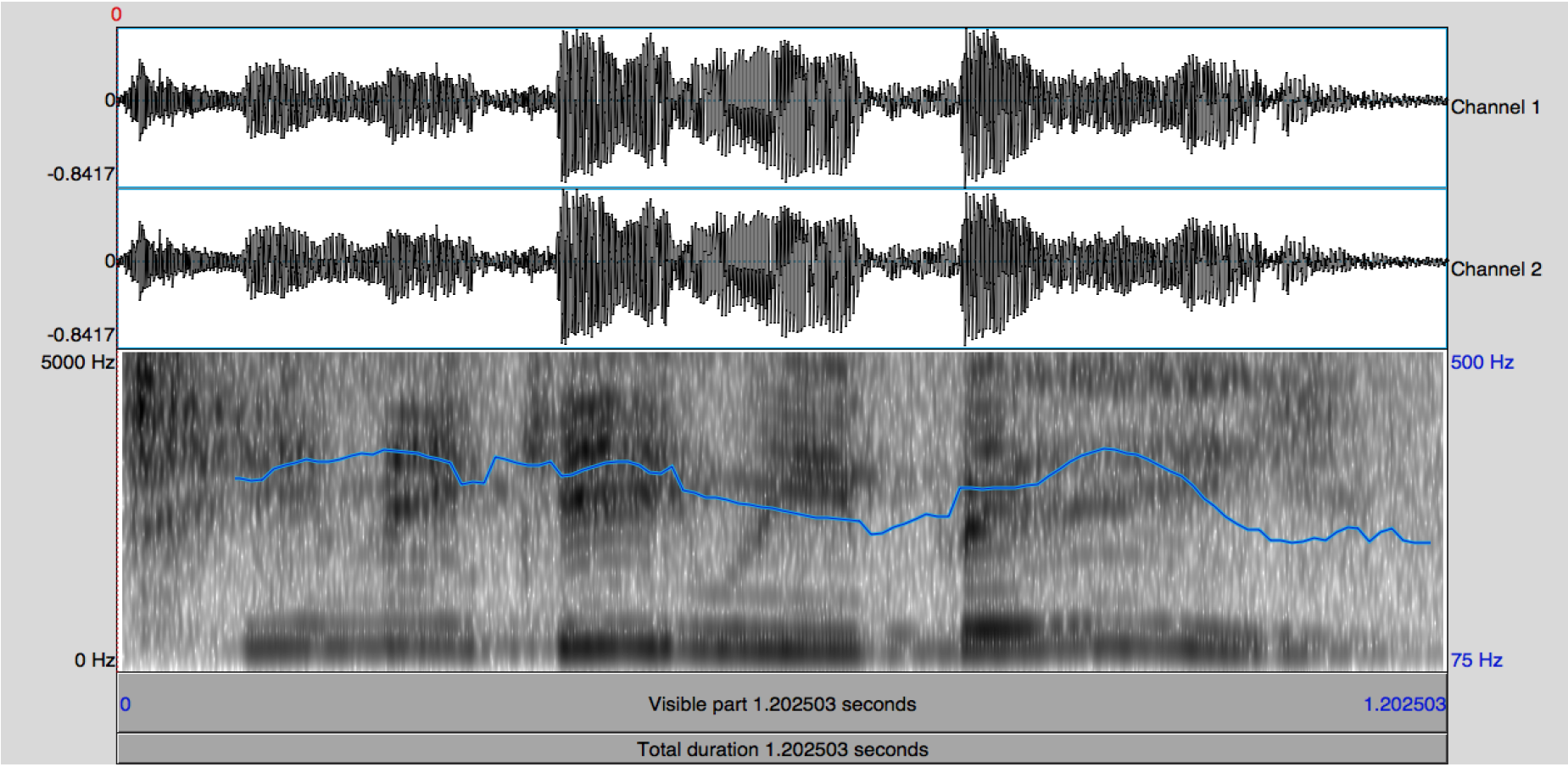

Figure E1: Praat Screenshot of the expression of Fearful vocalised on "chee nee gee mee bee" by a native speaker of English. (Encoding experiment)

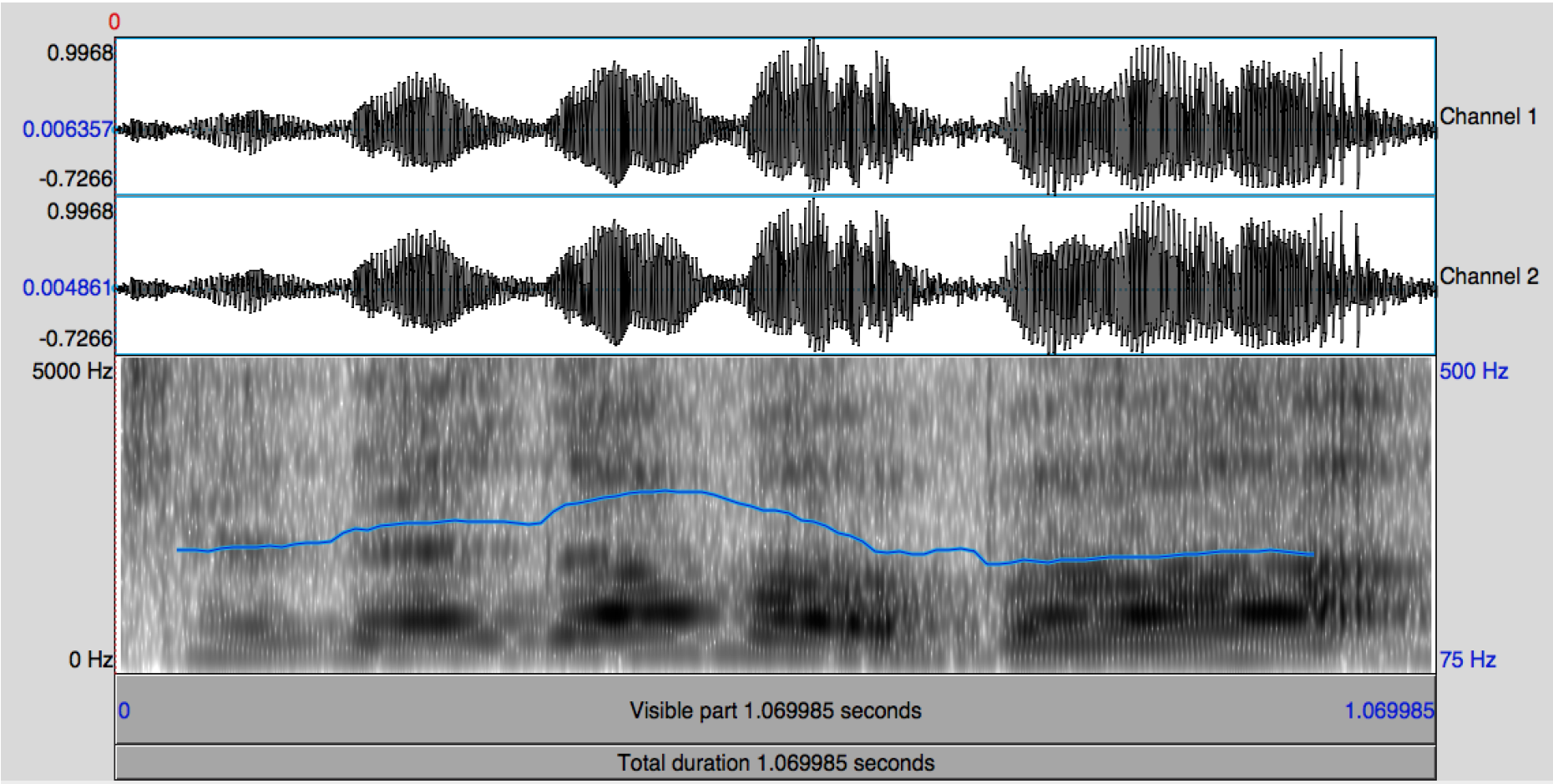

Figure E2: Praat Screenshot of the expression of Happy vocalised on “cha na ga ma ba” by a native speaker of Japanese. (Encoding experiment)

Appendix F: Acoustic analysis statistics

ENGLISH	Overall Significance rating	Pairwise emotion significance according to Gabriel	Ranked means
Sp rate (sylls per sec)	F(4,16) = 2.461, (*)p=0.87 (Levene p = 0.057) (Welch F(4, 6.062) = 4.118, p = 0.060)	(*)Happy/ Sad, p = 0.09.	HFCAS Fastest to slowest
Max Fo (Hz)	**F(4,16) = 11.762, ***p = 0.000 (Levene p = 0.058) (Welch F(4, 6.053) = 12.333, p = 0.005)**	***Sad/Happy, p = 0.001. **Sad/Angry, p = 0.008. (*)Sad/Fearful, p = .087. ***Calm/Happy, p = 0.001. **Calm/Angry, p = 0.007. (*)Calm/Fearful, p = 0.079	HAFSC Highest to lowest
Min Fo (Hz)	F(4, 16) = 3.839, *p = 0.023	*Fearful/Calm, p = 0.03. (*)Fearful/Angry, p = 0.082	CASHF Lowest to highest
Range Fo (Hz)	**F(4, 16) = 5.782, **p = 0.004**	*Angry/Sad, p = 0.016. *Angry/Fearful, p = 0.016. (*)Angry/Calm, p = 0.071. (*)Happy/Sad, p = .096. (*)Happy/Fearful, p = 0.088.	AHCSF Widest to narrowest
Mean Fo (Hz)	**F(4, 16) = 9.242, ***p = 0.000**	*Happy/Sad, p = 0.015. *Happy/Calm p= 0.019. *Angry/Sad, 0.014. *Angry/Calm, p = 0.017. **Fearful/Sad, p = 0.007. **Fearful/Calm, p = 0.009	FAHCS Highest to lowest
Fo SD (Hz)	No significant distinctions		AHCSF Highest to lowest
90th Perc f0 (Hz)	**F(4, 16) = 10.859, ***p = 0.000**	***Angry/Sad, p = 0.00. ***Angry/Calm, p = 0.00. (*)Angry/Happy, p = 0.06. *Fearful/Sad, p = 0 .035. *Fearful/Calm, p = 0.037.	AFHCS Highest to lowest
10th Perc Fo (Hz)	No significant distinctions		HFSCA Lowest to highest
80 Perc Range Fo (Hz)	F(4, 16) = 4.233, *p = 0.016	*Sad/Angry, p = 0.029. (*)Sad/Fearful, p = 0.091.	AFHCS Highest to lowest
Mean Int (dB)	F(4, 16) = 3.813, *p = 0.023	(*)Sad/Happy, p = 0.073 *Sad/Angry, p = 0.024 *Sad/Fearful, p = 0.034	AFHCS Highest to lowest.

Range Int (dB)	No significant distinctions	AHFCS Highest to lowest.

***p< 0.001 **p<0.01 *p<0.05 (*)p<0.099. (*) indicates values which are tending towards a significant level.

Table F1: Significant distinctions between emotions by acoustic parameter for English using Gabriel post-hoc test. H=Happy; S=Sad; A=Angry; F=Fearful; C=Calm.

JAPANESE	Overall Significance rating	Pairwise emotion significance according to Gabriel	Ranked means
Sp rate (sylls per sec)	F(4,16) = 0.818, p = 0.532	No significant distinctions	HCAFS Fastest to slowest.
Max Fo (Hz)	**F(4,6.698) = 11.762, **p = 0.004 (Levene ***p = 0.000)**	(*)Sad/Happy, p = 0.052. **Sad/Fearful, p = 0.002. *Calm/Happy, p = 0.05. **Calm/Fearful, p = 0.003	FHASC Highest to lowest.
Min Fo (Hz)	No significant distinctions		FACSH Lowest to highest.
Range Fo (Hz)	**F(4, 7.609) = 202.176, ***p = 0.000 (Levene ***p = 0.000)**	***Angry/Sad, p=0.001.**Angry/Fearful, p=0.008.*Angry/Calm, p=0.025.*Happy/Sad, p=0.04.***Fearful/Happy, p=0.000.***Fearful/Sad, p=0.000. ***Fearful/Calm, p=0.000	FAHCS Widest to narrowest.
Mean Fo (Hz)	F(4, 6.718) = 4.015, (*)p = 0.056. (Levene *p = 0.014)	(*)Happy/Sad, p = 0.074.*Happy/Calm p= 0.027.	HFASC Lowest to highest
Fo SD (Hz)	**F(4, 6.481) = 20.049, ***p = 0.001. (Levene: *p = 0.034)**	***Fearful/Sad, p = 0.000. **Fearful/Calm, p = 0.004.	FAHCS Highest to lowest.
90thPercf0 (Hz)	**F(4, 16) = 5.747, **p = 0.005**	*Angry/Sad, p = 0.038.**Angry/Calm, p = 0.004.	AFHSC Highest to lowest
10thPercf0 (Hz)	No significant distinctions		CSAFH Lowest to highest
80% range Fo (Hz)	**F(4, 16) = 6.110, **p = 0.004**	(*)Angry/Happy, p = 0.094.**Angry/Sad, p=0.007.**Angry/Calm, p=0.006.	AFHSC Widest to narrowest.
Mean Int (dB)	No significant distinctions		AFHSC Highest to lowest.
Range Int (dB)	No significant distinctions		SHAFC Highest to lowest.

***p< 0.001 **p<0.01 *p<0.05 (*)p<0.099. (*) indicates values which are tending towards a significant level.

Table F2: Significant distinctions between emotions by acoustic parameter for Japanese using Gabriel post-hoc test. H=Happy; S=Sad; A=Angry; F=Fearful; C=Calm.

ENGLISH AND JAPANESE	Happy	Sad	Angry	Fearful	Calm	Overall
Sp rate (sylls per sec)	none	F(1,8) = 5.214, **(*)p = 0.052**	none	none	F(1,8) = 6.409, ***p = 0.035**	F(1,40) = 7.131, ***p = 0.011**
Max f0 (Hz)	F(1,7) = 5.211, **(*)p = 0.056**	F(1,8) = 8.137, ***p = 0.021**	F(1,5) = 10.222, ***p = 0.024**	none	none	none F(1,32.263) = 0.322, p = 0.574 (Levene ***p = 0.001)
Min f0 (Hz)	none	none	none	F(1,5) = 6.170, **(*)p = 0.058**	none	none F(1,31.804) = 2.043, p = 0.161 (Levene **p = 0.003)
Range f0 (Hz)	F(1,7) = 5.082, **(*)p = 0.059**	none	F(1,5) = 15.541, ***p = 0.011**	F(1,5) = 57.438, ****p = 0.001**	F(1,6.703) = 4.399, **(*)p = 0.076** (Levene *p = 0.038)	none F(1,40) = 2.680, p = 0.109
Mean f0 (Hz)	none	F(1,5.188) = 8.566, ***p = 0.031**	none	none	none	none F(1,30.040) = 0.335, p = 0.567 (Levene ***p = 0.001)
SD f0 (Hz)	none	none	F(1,2.212) = 8.214, **(*)p = 0.092** (Levene *p = 0.026)	F(1,5) = 14.691, ***p = 0.012**	none	none F(1,34.622) = 1.016, p = 0.320 (Levene *p = 0.045)
90 Perc f0 (Hz)	none	F(1,8) = 4.865, **(*)p = 0.058**	none	none	none	none F(1,40) = 0.622, p = 0.435
10 Perc f0 (Hz)	none	None	none	none	F(1,8) = 25.150, ****p = 0.001**	none F(1,40) = 2.335, p = 0.134
80Perc Range f0(Hz)	none	none	none	none	none	none F(1,40) = 1.557, p = 0.219
Mean Int (dB)	none	none	none	none	none	none F(1,40) = 0.007, p = 0.932

RangeInt (dB)	none	none	none	none	F(1,8) = 3.560, (***)p = 0.096**	none F(1,40) = 0.441, p = 0.510

***p< 0.001 **p<0.01 *p<0.05 (*)p<0.099. (*) indicates values which are tending towards a significant level.

Table F3: Significant distinctions between English and Japanese in the mean level of each acoustic variable for each emotion and overall. H=Happy; S=Sad; A=Angry; F=Fearful; C=Calm.

References

Abelin, Å. (2004) Cross-cultural multimodal interpretation of emotional expressions – an experimental study of Spanish and Swedish. *Proceedings of Speech prosody 2004*, Nara, Japan, 647–650.

Abelin, A., & Allwood, J. (2000). Cross-linguistic interpretation of emotional prosody. In: R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.) (2000). *Proceedings of the ISCA Workshop on Speech and Emotion*, Northern Ireland, September 5-7, 110-113.

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh University Press.

Adler, S. (1988). The technique of acting. Bantam Books.

Albas, D. C., McCluskey, K. W., Albas, C. A. (1976). Perception of the emotional content of speech: A comparison of two Canadian groups. Journal of Cross-Cultural Psychology, 7(4), 481-490.

Anolli, L., Wang, L., Mantovani, F., & De Toni, A. (2008). The voice of emotion in Chinese and Italian young adults. Journal of Cross-Cultural Psychology, 39, 565-598.

Aristotle (1991). The art of rhetoric. (H. C. Lawson-Tancred, Trans.). London: Penguin. (Original work 4th century BC)

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

Auberge, V., Audibert, N. & Rilliard, A. (2004.) Acoustic morphology of expressive speech: What about contours? In *Proceedings of Speech* prosody *International Conference*, March 23-26, 2004. Nara, Japan.

Audibert, N., Aubergé, V., & Rilliard, A. (2008). How we are not equally competent for discriminating acted from spontaneous expressive speech. In *Proceedings of Speech* prosody *International Conference*, May 6-9, 2008, Campinas, Brazil.

Audibert, N., Aubergé, V., & Rilliard, A. (2010). Prosodic correlates of acted vs. spontaneous discrimination of expressive speech: a pilot study. In *Proceedings of Speech prosody International Conference*, May 10-14, *2010*, Chicago, IL, USA. Retrieved from http://www.isca-speech.org/archive.

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

Bachorowski, J. A. & Owren, M. J. (2003). Sounds of emotion production and perception of affect-related vocal acoustics. *Annals of the New York Academy of Sciences, 1000,* 244–265. Retrieved from http://onlinelibrary.wiley.com

Balkwill, L. L., & Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception: An Interdisciplinary Journal*. *17*(1), 43-64. University of California Press Article Stable URL: http://www.jstor.org/stable/40285811

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

Banziger, T. & Scherer, K.R. (2005). The role of intonation in emotional expressions. *Speech Communication 46,* 252–267

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., & Aharonson, V. (2007). The impact of f0 extraction errors on the classification of prominence and emotion. In *Proceedings of ICPhS 2007*, Saarbrucken, Germany, 2201–2204.

Beier, E. G., & Zautra, A. J. (1972). Identification of vocal communication of emotions across cultures. *Journal of Consulting and Clinical Psychology, 39*, 166

Boersma, P. (2004) 23:09 +02.00 16-6-2004. praat-users group.

Boersma, P. (2005) 23:22 +02.00 29-7-2005. praat users group.

Boersma, P., & Weenink, D. (2007). Praat: doing phonetics by computer. [Computer program]. Version 4.6, retrieved 6 March 2007 from http://www.praat.org/

Bouchard, T. J. (1994). Genes, environment, and personality. *Science, 264*, 1700-1701.

Braun, A. & Oba, R. (2007). Speaking tempo in emotional speech – a cross-cultural study using dubbed speech. *Proceedings of the International Workshop on Paralinguistic Speech*, 3 August 2007, Saarbrücken, Germany.

Breitenstein, C.; Van Lancker, D.; Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and emotion, 15* (1), 57-80.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments.* Berkeley: University of California Press.

Burkhardt, F., Audibert, N., Malatesta, L., Turk, O., Arslan, L. & Auberge, V. (2006). Emotional Prosody - Does Culture Make A Difference? In R. Hoffmann & H. Mixdorff (Eds) *Proceedings of Speech prosody International Conference* May 2-5, 2006, Dresden, Germany.

Cannon, W. B. (1927). The James-Lange theory of emotion: A critical examination and an alternative theory. *American Journal of Psychology, 39,* 10-124.

Carr, L., Iacoboni, M., Dubeau, M. C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences of the United States of America, 100*, 5497–5502.

Catford, J. C. (1977). *Fundamental problems in phonetics.* Edinburgh University Press.

Chung, S., (2000). *Expression and perception of emotion extracted from the spontaneous speech in Korean and English*. Doctoral Dissertation, ILPGA, Sorbonne Nouvelle University, Paris, France.

Clark, D. M. (1983). On the induction of depressed mood in the laboratory: Evaluation and comparison of the Velten and musical procedures. *Advances in Behavioral Research and Therapy, 5,* 27–49.

Cowie, R. (2009). Perceiving emotion: towards a realistic understanding of the task *Phil. Trans. R. Soc. B 364*, 3515-3525.

Cowie, R. (2011). Editorial: 'Ethics and Good Practice' – Computers and Forbidden Places: Where Machines May and May Not Go. In Petta, P., Cowie, R. & Pelachaud, C. (Eds.) (2011). *Emotion-oriented systems: The Humaine Handbook (Cognitive Technologies)*, 707-711. Berlin, Heidelberg: Springer-Verlag.

Cowie, R., Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication, 40,* 5–32.

Cowie, R., Douglas-Cowie, E., Appolloni, B., Taylor, J., Romano, A., & Fellenz, W. (1999). What a neural net needs to know about emotion words. In N. Mastorakis, (Ed.), *Computational Intelligence and Applications,* 109–114. World Scientific & Engineering Society Press.

Cowie R., Douglas-Cowie, E., & Cox, C., (2005). Beyond emotion archetypes: databases for emotion modelling using neural networks. *Neural Networks 18*(4), 371-388.

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

Cutler, A., & Otake, T. (1999). Pitch accent in spoken-word recognition in Japanese. Journal of the Acoustical Society of America, 105, 1877–1888.

Dang, J., Li, A., Erickson, D., Suemitsu, A., Akagi, M., Sakuraba, K., Minematsu, N. & Hirose, K. (2009). Comparison of Emotion Perception among Different Cultures. *Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference*, Sapporo, Japan, 538-544.

Darwin, C. (1998). The expression of the emotions in man and animals (3rd ed.). P. Ekman (Ed.). London: Harper Collins; New York: Oxford University Press. (Original work published 1872)

Davitz, J. R. (1964). *The Communication of Emotional Meaning.* New York: McGraw-Hill Book Co.

Denes, P. B. & Pinson, E. N. (1993). *The Speech Chain: The physics and biology of spoken speech.* (2nd ed.) New York: W.H. Freeman and Company.

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

Douglas-Cowie, E., Cowie, R., Schroeder, M., (2000). A new emotion database: Considerations, sources and scope. In: R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.) (2000). *Proceedings of the ISCA Workshop on Speech and Emotion,* 5-7 September, 2000, Newcastle, Northern Ireland.

Douglas-Cowie, E., Devillers, L., Martin, J-C. Cowie, R, Savvidou, S., Abrilian, S., & Cox, C (2005). Multimodal databases of everyday emotion: Facing up to complexity. *Proceedings of Interspeech,* September 4-8, 2005, Lisbon, Portugal.

Dromey, C., Silveira, J., & Sandor, P. (2005). Recognition of affective prosody by speakers of English as a first or foreign language. *Speech Communication*, *47*(3), 351-359.

Ekman P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole. (Ed.), *Nebraska symposium on motivation 1971* (pp.207–282). Lincoln: University of Nebraska Press.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion, 6* (3/4), 169-200.

Ekman, P. (1999). Basic emotions. In T. Dalgleish and T. Power (Eds.), *The handbook of cognition and emotion,* New York: John Wiley.

Ekman, P. & Friesen, W.V., (1971), Constants across cultures in the face and emotion, *Journal of Personality and Social Psychology, 17*(2), 124–129.

Ekman, P. & Friesen, W. V. (1975). *Unmasking the face. A guide to recognizing emotions from facial clues*. Englewood Cliffs, New Jersey: Prentice-Hall.

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Davidson, R.J., Ricard, M. & Wallace, B.A. (2005). *Buddhist and Psychological Perspectives on Emotions and Well-Being, 14* (2) American Psychological Society.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: guide-lines for research and an integration of findings.* New York: Pergamon Press.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes between emotions. *Science, 221*(4616), 1208-1210.

Ekman, P., Sorenson, E. R., & Friesen. W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science, New Series, 164*(3875), 86-88.

Eldred, S.H., & Price, D.B. (1958). The linguistic evaluation of feeling states in psychotherapy. *Psychiatry, 21,* 115-121.

Elfenbein, A., & N. Ambady, (2002). Is there an In-group advantage for emotion recognition? *Psychological Bulletin, 128(2),* 243-249.

Elfenbein, H.A., & Ambady, N. (2003). Universals and Cultural Differences in Recognizing Emotions. *Current Directions in Psychological Science, 12*(5), 159-164.

Ellsworth, P.C., & Scherer, K.R. (2003). Appraisal processes in emotion. In R.J. Davidson, K.R., Scherer, & H.H. Goldsmith (Eds.), Handbook of affective sciences, 572-595. New York: Oxford University Press.

'emotion'. In Oxford Dictionaries online. (2015). Retrieved from http://www.oxforddictionaries.com/definition/english/emotion

Erickson, D. (2006). Some gender and cultural differences in perception of vocally-expressed affect. In R. Hoffmann & H. Mixdorff (Eds) *Proceedings of Speech prosody International Conference* May 2-5, 2006, Dresden, Germany.

Erickson, D. (2010). Perception by Japanese, Korean and American listeners to a Korean speaker's recollection of past emotional events: Some acoustic cues. *Proceedings of Speech prosody International Conference 2010*, Chicago, USA.

Erickson, D., Huang, C-F., Shochi, T., Rilliard, A., Dang, J., Iwata, R., & Lu, X. (2008a). Acoustic and articulatory cues for Taiwanese, Japanese and American listeners' perception of Chinese happy and sad speech. *Proceedings of the 2008 Autumn Meeting of The Acoustical Society of Japan*, 351-354.

Erickson, D., Rilliard, A., Shochi, T., Han, J., Kawahara, H., & Sakakibara, K. (2008b). A cross-linguistic-comparison of perception to formant frequency cues in emotional speech. *COCOSDA,* Kyoto, Japan, 163-167.

Ethofer, T., Erb, M., Anders, S., Wiethoff, S., Herbert, C., Saur, R., Grodd, W., & Wildgruber, D. (2006). Effects of prosodic emotional intensity on activation of associative auditory cortex. *NeuroReport, 17,* 249–253.

Ethofer, T., Van De Ville, D, Scherer, K. & Vuilleumier, P. (2009). Decoding of emotional information in voice-sensitive cortices. *Current Biology 19*(12)*,* 1028–1033.

Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General, 113*, 464-486.

Ferguson, C. A. (1978). Historical background of universals research. In J.H. Greenberg, C.A. Ferguson, E.A. Moravcsik (Eds.) *Universals of human language, Vol. One: Method and theory*, 7–33. Stanford, CA: Stanford University Press.

Field, A. (2009). *Discovering statistics using SPSS*. London: Sage.

Fonagy, I. & Magdics, K. (1963). Emotional patterns in intonation and music. *Zeitschrift fur Phonetik 16*, 293-326.

Fonagy, I., (1981). Emotions, voice and music. In J. Sundberg (Ed.), *Research Aspects On Singing: Proceedings From A Seminar Organised By The Committee For The Acoustics Of Music, 33,* 51–79. Royal Swedish Academy of Music.

Frick, R. W. (1985). Communicating emotion: the role of prosodic features. *Psychological Bulletin. 97*, 412–429.

Friesen, W. V. (1972). *Cultural differences in facial expressions in a social situation: An experimental test of the concept of display rules*. (Unpublished doctoral dissertation). University of California, San Francisco.

Galatà, V., & Romito, L. (2010). Un corpus sperimentale per lo studio cross-linguistico europeo delle emozioni vocali. In: *Proceedings of the 5th AISV Conference - La dimensione Temporale del parlato,* University of Zürich, Switzerland, February 4-6 2009, 603-641.

Gerrig, R. J., & Zimbardo, P. G. (2002). Glossary of psychological terms. In *Psychology and life* (16th ed.). Boston, MA: Allyn and Bacon, Pearson Education.

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

Graham, C. R., Hamblin, A. W., & Feldstein, S. (2001). Recognition of emotion in English voices by speakers of Japanese, Spanish and English. *International Review of Applied Linguistics in Language Teaching,* 39(1), 19-37.

Greasley, P., Sherrard, C., & Waterman M. (2000). Emotion in language and speech: Methodological issues in naturalistic approaches. *Language and Speech, 43*, 355–375.

Hatfield, E., Cacioppo, J., & Rapson, R. L. (1994). *Emotional contagion.* New York: Cambridge University Press.

Hayashi, Y. (1999). Recognition of vocal expression of emotions in Japanese: using the interjection "*eh".* In *Proceedings of the 14th International Conference of Phonetic Sciences*, Univ. of California, Berkeley.

Hennenlotter, A., Dresel, C., Castrop, F., Ceballos-Baumann, A. O., Wohlschlager, A. M., & Haslinger, B. (2008). The link between facial feedback and neural activity within central circuitries of emotion - New insights from botulinum toxin-induced denervation of frown muscles. *Cerebral Cortex (2009), 19*(3), 57-52.

Heylen, D., Bevacqua, E., Pelachaud, C., Poggi, I., Gratch, J., & Schröder, M. (2011). Generating listening behaviour. In P. Peta, C. Pelachaud, & R. Cowie (Eds.), *Emotion-Oriented Systems: The HUMAINE Handbook,* 321-348. Berlin: Springer Verlag.

Ibrakhim, I. (2004). Universal and linguistic features of expressing emotional information: Differentiation in the perception level. In *Speech Prosody 2004,* Nara, Japan. March 23-26. Retrieved from http://www.isca-speech.org/archive

Ishii, K., Reyes, J. A., & Kitayama, S. (2003). Spontaneous attention to word content versus emotional tone: Differences among three cultures. *Psychological Science, 14*(1), 39-46.

Izard, C. E. (1971). *The face of emotion*. New York: Appleton-Century-Crofts.

Izard, C. E. (1994). Innate and Universal Facial Expressions: Evidence From Developmental and Cross-Cultural Research. *Psychological bulletin, 115*(2), 288-299.

James, W. (1983). *The Principles of Psychology*. Cambridge, MA: Harvard University Press. New York: H.Holt and Company. (First published in 1890 by New York: H.Holt and Company).

Jaywant, A., & Pell, M.D. (2012). Categorical processing of negative emotions from speech prosody. *Speech Communication 54*(1), 1–10.

Johnstone, T., Van Reekum, C.M., Scherer, K.R. (2001). Vocal correlates of appraisal processes. In K.R. Scherer, A. Schorr & T. Johnstone (Eds) *Appraisal processes in emotion: Theory, Methods, Research.* 271-284. New York: Oxford University Press.

Jurgens, U. & Hammerschmidt K. (2006). Common acoustic features in the vocal expression of emotions in monkeys and man. *Primate Report, 74*, 3-8.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on decoding accuracy and cue utilization in vocal expression of emotion. *Emotion, 1*, 381–412.

Juslin, P. & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*, 770-814.

Juslin, P., & Sloboda, J. (Eds.) (2001). *Music and emotion: Theory and research.* USA: Oxford University Press.

Keating, P. & G. Kuo, (2010) Comparison of speaking fundamental frequency in English and Mandarin. *UCLA Working Papers in Phonetics, No. 108*, pp. 164-187.

Kenealy, P. (1988). Validation of a music mood induction procedure: Some preliminary findings. *Cognition and Emotion, 2,* 41–48.

Kilbride, J.E., & Yarczower, M. (1983). Ethnic bias in the recognition of facial expressions. *Journal of Nonverbal Behavior, 8*, 27-41.

Kitayama, S., & Ishii, K. (2002). Word and voice: Spontaneous attention to emotional speech in two cultures. *Cognition and Emotion, 16,* 29–59.

Kotz, S.A. & Paulmann, S. (2007). When Emotional Prosody and Semantics Dance Cheek to Cheek: ERP Evidence. *Brain Research 1151*, 107–118.

Kramer, E. (1964). Elimination of verbal cues from judgements of emotion from voice. *Journal of Abnormal and Social Psychology, 68,* 390-396.

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. *NeuroImage, 37,* 1445–1456.

Krumhuber, E.G. & Scherer, K.R. (2011). Affect bursts: Dynamic patterns of facial expression. *Emotion, 11*(4), 825-841.

Ladefoged, P. (1975). *A course in phonetics.* New York: Harcourt, Brace, Jovanovich.

Lange, C. G. (1885) The emotions. In C. G. Lange and W. James (Eds.) (1967), *The emotions.* New York: Hafner. [facsimile of 1922 ed.]

Laukka, P., Audibert, N. & Auberge, V., (2007). Graded structure in vocal expression of emotion: What is meant by "prototypical expressions"? *Proceedings of 1st International Workshop on Paralinguistic Speech*, Saarbrucken, Germany, 1-4.

Laukka, P., Elfenbein, H. A., Chui, W., Thingujam, N. S., Iraki, F. K., Rockstuhl, T., & Althoff, J. (2010). Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. In L. Devillers, B. Schuller, R. Cowie, E. Douglas-Cowie, & A. Batliner (Eds.), *Proceedings of the LREC 2010 Workshop on Corpora for Research on Emotion and Affect,* 53-57. Paris, France: European Language Resources Association.

Laver, J. (1981). *Users' manual for vocal profile analysis protocol: A perceptual guide.* Unpublished paper, University of Edinburgh.

Laver, J. D., Wirz, S. L., Mackenzie, J., & Hiller, S. (1981). A perceptual protocol for the analysis of vocal profiles. *University of Edinburgh Department of Linguistics Work in Progress, 14*, 139-155.

Le Doux, J.E. (1996). *The Emotional Brain.* New York: Simon and Schuster.

Lee, T. W., Josephs, O., Dolan, R. J., & Critchley, H. D. (2006). Imitating expressions: emotion-specific neural substrates in facial mimicry. *Social Cognitive and Affective Neuroscience, 1*(2), 122–135.

Leys, R. (2010). How did fear become a scientific object and what kind of object is it? *Representations*, *110*(1), 66-104.

Livingstone, S.R., Peck, K. & Russo, F.A. (2013). Acoustic differences in the speaking and singing voice. *Musical Acoustics Session 5aMUb: Digital Libraries for Speech and Singing. I*CA 2013 Montreal, Canada.

Matsumoto, D. (1996). *Unmasking Japan.* Stanford, CA: Stanford University Press,

Matsumoto, D. (2006). Culture and nonverbal Behavior. In *The Sage Handbook of Nonverbal Communication*. Thousand Oaks, CA: Sage Publications Inc.

McCluskey, K. W., Albas, D. C., Niemi, R. R., Cuevas, C., & Ferrer, C. A. (1975). Cross-cultural differences in the perception of the emotional content of speech: A study of the development of sensitivity in Canadian and Mexican children. *Developmental Psychology 11*(5), 551-555.

McCluskey, K. W., & Albas, D. C. (1981). Perception of the emotional content of speech by Canadian and Mexican children, adolescents, and adults. *International Journal of Psychology 16*, 119–132.

Mead, M. (1928). *Coming of Age in Samoa*. New York: Harper Collins.

Menezes, C., Erickson, D., & Franks, C. (2010). Comparison between linguistic and affective perception of sad and happy – a cross-linguistic study. In *Speech Prosody 2010*, Chicago, IL, USA, May 10-14.

Mennen, I., Schaeffler, F. & Docherty, G. (2008). Cross-language differences in pitch range: Full Research Report ESRC End of Award Report, RES-000-22-1858. Swindon: ESRC.

Mennen, I., Schaeffler, F. & Docherty, G. (2012). Cross-language difference in f0 range: a comparative study of English and German. *Journal of the Acoustical Society of America*, 131(3), 2249-2260.

Mesquita, B. & Karasawa (2002). Different Emotional Language. *Cognition and Emotion, 16(1),* 127-141

Mesquita, B., & Walker, R. (2003). Cultural differences in emotions: a context for interpreting emotional experiences. *Behaviour, Research and Therapy, 41*, 777-793.

Messinger, D. S., Fogel, A., & Dickson, K. (2001). All smiles are positive, but some smiles are more positive than others. *Developmental Psychology, 37*(5), 642-653.

Mozziconacci, S. J. L. (1998). *Speech variability and emotion: Production and perception.* (Ph.D. thesis) Eindhoven, The Netherlands.

Mozziconacci, S. (2002). Prosody and emotions. *Proceedings of the 1st International Conference on Speech Prosody*, Aix-en-Provence, France, April 11-13. 1-9.

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

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

Murray, I. R., & Arnott, J. L., (2008). Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. *Computer Speech and Language, 22*(2), 107-129.

Nakamichi, A., Jogan, A., Usami, M., & Erickson, D. (2002). Perception by native and non-native listeners of vocal emotion in a bilingual movie. *Gifu City Women's College Research Bulletin*, *52*, 87–91.

Ogarkova, A., Borgeaud, P., & Scherer, K. R. (2009). Language and culture in emotion research: A multidisciplinary perspective. *Social Science Information 48*, 339-357.

Ohala, J.J. (1984). An ethological perspective on common cross-language utilization of f0 of voice. *Phonetica, 41*, 1-16.

Ohata, K. (2004). Phonological differences between Japanese and English: Several potentially problematic areas of pronunciation for Japanese ESL/EFL Learners. *Asian EFL Journal, 6*(4), Article 5.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language, 32*(2)*,* 258-278.

Owren, M. J., & Bachorowski, J. A. (2007). Measuring emotion-related vocal acoustics. In J. Coan, & J. Allen (Eds.), *Handbook of emotion elicitation and assessment, 239-266.* Oxford: Oxford University Press.

Panksepp, J. (1998). *Affective Neuroscience: The foundations of human and animal emotions.* New York: Oxford University Press.

Papousek, H., Jurgens, U., & Papousek, M. (Eds.). (1992). *Nonverbal vocal communication: Comparative and developmental approaches.* Cambridge, England: Cambridge University Press.

Paulmann, S., & Pell, M. D. (2010). Contextual influences of emotional speech prosody on face processing: How much is enough? *Cognitive, Affective and Behavioral Neuroscience, 10,* 230–242.

Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *Journal of the Acoustical Society of America, 109*(4), 1668–1680.

Pell, M. D. (2002). Evaluation of nonverbal emotion in face and voice: Some preliminary findings on a new battery of tests. *Brain and Cognition, 48,* 499–504.

Pell, M. D. (2005a). Nonverbal emotion priming: Evidence from the 'facial affect decision task'. *Journal of Nonverbal Behavior, 29*(1), 45–73.

Pell, M. D. (2005b). Prosody-face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior, 29*(4), 193–215.

Pell, M. D., & Kotz, S. A. (2011). On the time course of vocal emotion recognition. *PLoS One, 6*(11).

Pell, M. D., Kotz, S. A., Paulmann, S., & Alasseri, A. (2005). Recognition of basic emotions from speech prosody as a function of language and sex. *Abstracts of the Psychonomic Society 46th Annual Meeting*, November 10-13. Toronto, Canada, Poster 3070.

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009a). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior, 33(*2), 107-120.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009b). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics, 37(4)*, 417-435.

Pell, M. D., & Skorup, V. (2008). Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication, 50,* 519–530.

Pellegrino, F., Coupé C. & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, *87* (3).

Petta, P., Cowie, R. & Pelachaud, C. (Eds.) (2011). Emotion-oriented systems: The Humaine Handbook (Cognitive Technologies). Berlin, Heidelberg: Springer-Verlag.

Pfitzinger,H.R., Mixdorff, H. & Schwarz, J. (2009) Comparison of Fujisaki-model extractors and F0 stylizers. *Proceedings of Interspeech*, 2009, 2455–2458.

Pfitzinger, H. R., Amir, N., Mixdorff, H. & Bösel, J. (2011). Cross-language perception of Hebrew and German authentic emotional speech. *Proceedings of ICPhS XVII. Regular Session.* Hong Kong, August 17-21, 1586-1589.

Pike, K. L. (1946). *The Intonation of American English.* University of Michigan Press.

Pinchot-Kastner, M., & Crowder, R. (1990). Perception of the major/minor distinction: emotional connotation in young children. *Music Perception, 8*(2), 189–202.

Posamentier, M., & Abdi, H. (2003). Processing faces and facial expressions. *Neuropsychology Review, 13*(3), 113-144.

Prinz, J. J. (2004). Which emotions are basic? In D. Evans & P. Cruse (Eds.), *Emotion, evolution, and rationality*. Oxford University Press.

Radulovic, J. & Stankovic, B. (2007) Genetic Determinants of Emotional Behavior: Legal Lessons from Genetic Models (March 1, 2007). *DePaul Law Review*, Vol. 56, 823-836.

Reisenzein, R. (1983). The Schachter theory of emotion: Two decades later. Psychological Bulletin, 94(2), 239-264.

Rigoulot, S., & Pell, M. D. (2012). Seeing emotion with your ears: Emotional prosody implicitly guides visual attention to faces. PLoS One, 7(1).

Rilliard, A., Shochi, T., Martin, J.C., Erickson, D. and Aubergé, V. (2009). Multimodal Indices To Japanese And French Prosodically Expressed Social Affects. *Language and Speech 52* (2&3) 223- 243.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S. and Setter, J. (1998) Transcription of Prosodic and Paralinguistic Features of Emotional Speech, *Journal of the International Phonetic Association*, 28, 83-94.

Sadfar, S., Friedlmeier, W., Matsumoto, D., Yoo, S. H., Kwantes, C. T., Kakai, H., & Shigemasu, E. (2009). Variations of emotional display rules within and across cultures: A comparison between Canada, USA and Japan. In *Canadian Journal of Behavioural Science, 41*(1), 1-10.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalisations. *Proceedings of the National Academy of Sciences, 107*(6)*,* 2408–2412.

Sauter, D.A. & Scott, S.K. (2007). More than one kind of happiness: can we recognize vocal expressions of different positive states? *Motivation and Emotion, 31*(3), 192-199.

Schachter, S., & Singer, J. (1962). Cognitive, social and physiological determinants of emotional state. *Psychology Review, 69*, 379-99.

Scheffers, M.T.M. (1988). Automatic stylization of FO-contours. In W.A. Ainsworth & J.N. Holmes (eds.): *Proceedings of the 7th FASE Symposium, Edinburgh, 3*, 981-987

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer, & P. Ekman (Eds.), *Approaches to Emotion*, 293-317. Hillsdale, NJ: Lawrence Erlbaum.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143-165.

Scherer K. R. (1989). *Handbook of social psychophysiology.* H. Wagner & A. Manstead (Eds.), 170 et seq. Wiley.

Scherer, K. R. (1994). Affect bursts. In S. H. M. van Goozen, N. E. van de Poll, & J. A. Sergeant (Eds.), *Emotions* (pp. 161-193). Hillsdale, NJ: Lawrence Erlbaum.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal Processes in Emotion: Theory, Methods, Research*, 92-120. Oxford University Press.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*, 227-256.

Scherer, K. R. (2004). Which emotions can be induced by music? What are the underlying Mechanisms? And how can we measure them? *Journal of New Music Research. 33*(3), 239-251.

Scherer, K. R. (2013a). The evolutionary origin of multimodal synchronization in emotional expression. *Journal of Anthropological Sciences, 91,* 185-200.

Scherer, K. R. (2013b). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech and Language* 27, 40–58.

Scherer, K. R., Banse, R, & Walbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of cross-cultural psychology*, 32 (1), 76-92.

Scherer, K. R., & Ceschi, G. (2000). 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., Clark-Polner, E., & Mortillaro, M. (2011). In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology Volume, 46*(6) 2011, 401-435.

Scherer, K. R., & Ellgring, H. (2007). Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion*, *7*(1)*,* 158–171.

Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America, 76*, 1346-1356.

Scherer, K. R., Zentner M. R., & Schacht A. (2002). Emotional states generated by music: An exploratory study of music experts. *Musicae Scientiae. The Journal of the European Society for the Cognitive Sciences of Music, 6*(1), 149-171.

Schröder, M. (2000). Experimental study of affect bursts. *Proceedings of ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland*, September 2000.

Schröder, M. (2001) Emotional speech synthesis: a review. In *Proceedings Eurospeech 2001, ISCA*, Bonn, Germany, 561–564.

Schröder, M., (2003). Experimental study of affect bursts. *Speech Communication, 40*(1), 99–116.

Schröder, M. (2004). Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. (PhD thesis) *PHONUS 7, Research Report of the Institute of Phonetics*, Saarbrucken, Saarland University.

Schröder, M., Heylen, D., & Poggi, I. (2006). Perception of Non-Verbal Emotional Listener Feedback. In *Proceedings Speech Prosody 2006*, May 2-5, Dresden, Germany.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987): Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52,* 1061–1086.

Shiota, M, Keltner D, Campos B, & Hertenstein M (2004). Positive emotion and the regulation of interpersonal relationships. In P. Phillipot, & R.Feldman (Eds.), *Emotion regulation*, 127–156. Mahwah, NJ: Erlbaum.

Skinner, R. E. (1935). A calibrated recording and analysis of the pitch, force, and quality of vocal tones expressing happiness and sadness. And a determination of the pitch and force of the subjective concepts of ordinary, soft, and loud tones. *Speech Monographs, 2*, 81-137.

Sobin, C., & Alpert, M. (1999). Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy, *Journal of Psycholinguistic Research, 23*(4), 347–365.

Soto, J.A., Levenson, R.W., Ebling, R. (2005). Cultures of Moderation and Emotional Experience, Behavior, and Physiology in Chinese and Mexican Americans. *Emotion*, 5 (2), 154-165.

Spackman, M., Brown, B., & Otto, S. (2009). Do emotions have distinct vocal profiles? A study of idiographic patterns of expression. *Cognition & Emotion, 23*(8), 1565-1588.

Stanislavski, C. (1967). *An actor prepares*. Penguin.

Steidl, S., Batliner, A., No¨ th, E., & Hornegger, A. (2008). Quantification of segmentation and f0 errors and their effect on emotion recognition. *Proceedings of*

the 11th International Conference on Text, Speech and Dialogue, TSD 2008, 525–534.

Strasberg, L. (1987). *A dream of passion: The development of the Method*. USA: Penguin.

Thompson, W., & Balkwill, L.-L. (2006). Decoding speech prosody in five languages. *Semiotica, 158,* 407-424.

Tomkins, S. S. (1962). *Affect imagery consciousness: Vol. I. The positive affects*. New York: Springer.

Tomkins, S. S. (1963). *Affect imagery consciousness: Vol. II. The negative affects.* New York: Springer.

Tooby, J. & Cosmides, L. (1990). Evolutionary psychology and the emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions* (1990) 91–115. New York: Guilford Press.

van Bezooijen, R. (1984). *Characteristics and recognisability of vocal expressions of emotion.* Dordrecht: Foris.

van Bezooijen, R., Otto, S. A., & Heenan, T. A. (1983). Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology, 14*(4), 387–406.

van der Zwaag, M. D., Westerink J. H. D. M., & van den Broek E. L. (2009). Deploying music characteristics for an affective music player. *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction, ACI2009*, September 10-12. Amsterdam, The Netherlands, 459-465. IEEE Computer Society. ISBN 978-1-4244-4800-5.

Velten, E. (1968). A laboratory task for induction of mood states. *Behaviour Research and Therapy, 6*(4), 473–482.

Ververidis, D., & Kotropoulos, C. (2003). A state of the art review on emotional speech databases. In *Proceedings 1st Richmedia Conference*, Lausanne, Switzerland, 109–119.

Ververidis, D., & Kotropulos, C. (2006). Emotional speech recognition: Resources, features, and methods", *Speech Communication, 48*, 1162-1181.

Wallbott, H. G., & Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology, 51*(4), 690–699.

Warner, N., & Arai, T. (2001). Japanese mora-timing: A review. *Phonetica, 58*(1-2).

Webster, G., & Weir, C. (2005). Emotional responses to music: Interactive effects of mode, texture, and tempo. *Motivation and Emotion, 29*(1), 19–39.

Wicker, B., Keysers C., Plailly J., Royet J. P., Gallese V., & Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. *Neuron, 40*, 655–64.

Wierzbicka, A. (1992). *Semantics, culture, and cognition: Universal human concepts in culture-specific configurations.* Oxford: Oxford University Press.

Wierzbicka, A. (2009). Language and metalanguage: Key issues in emotion research. *Emotion Review*, *1*(1), 3-14.

Wildgruber, D., Ackermann, H., Kreifelts, B., Ethofer, T., (2006). Cerebral processing of linguistic and emotional prosody: fMRI studies. *Prog. Brain Res. 156*, 249–268.

Williams, C. & Stevens, K. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, *52*, 1238-1250.

Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics, 38,* 1559–1579.

Wilting, J., Krahmer, E. & Swerts, M. (2006). Real vs. acted emotional speech. *Proceedings* of *Interspeech 2006, 9th International Conference on Spoken Language Processing,* Pittsburgh, PA. USA, September 17-21.

Woods, J. H. (2003). *The Yoga Sutras of Patanjali*. NewYork: Dover Publications Inc; Cambridge, Mass: Harvard University Press.

Wundt, W. (1897). *Outlines of Psychology*. Translated by C.H. Judd. Leipzig: Wilhelm Engelmann (Reprinted Bristol: Thoemmes, 1999).

Xu, Y., & Chuenwattanapranithi, S. (2007). Perceiving anger and joy in speech through the size code. *Proceedings of 16th International Congress of Phonetic Sciences*, 2105-2108.

Yamazawa, H., and Hollien, H. (1992). Speaking fundamental frequency patterns of Japanese women, *Phonetica 49,* 128-140.

Yanushevskaya, I. Gobl, C., & Ní Chasaide, A., **(**2005). Voice quality and *f*0 cues for affect expression: implications for synthesis. *Proceedings of the 8th International Conference on Spoken Language Processing, INTERSPEECH 2005*. Lisbon, 1849-1852.

Yanushevskaya, R, Ni Chasaide, A. & Gobl, C. (2011). Universal and language specific perception of affect from voice. *Proceedings of ICPhS XVII. Regular Session*. Hong Kong, August 17-21, 2208-2211.

Zillmer, E., Spiers, M., & Culbertson, W. C. (2008). *Principles of neuropsychology.* Belmont, CA, USA: Wadsworth.

Zinken, J., Knoll, M., Panksepp, J. (2008). Universality and diversity in the vocalisation of emotions. In K. Izdebski (Ed.), *Emotions in the human* voice, 185-202. San Diego, CA: Plural Publishing.