An investigation into vocal expressions of emotions: the roles of valence, culture, and acoustic factors

Disa Sauter

Thesis submitted for the degree of Doctor of Philosophy
University College London, September 20, 2006

ABSTRACT

This PhD is an investigation of vocal expressions of emotions, mainly focusing on non-verbal sounds such as laughter, cries and sighs. The research examines the roles of categorical and dimensional factors, the contributions of a number of acoustic cues, and the influence of culture. A series of studies established that naïve listeners can reliably identify non-verbal vocalisations of positive and negative emotions in forced-choice and rating tasks. Some evidence for underlying dimensions of arousal and valence is found, although each emotion had a discrete expression. The role of acoustic characteristics of the sounds is investigated experimentally and analytically. This work shows that the cues used to identify different emotions vary, although pitch and pitch variation play a central role. The cues used to identify emotions in non-verbal vocalisations differ from the cues used when comprehending speech. An additional set of studies using stimuli consisting of emotional speech demonstrates that these sounds can also be reliably identified, and rely on similar acoustic cues. A series of studies with a pre-literate Namibian tribe shows that non-verbal vocalisations can be recognized across cultures. An fMRI study carried out to investigate the neural processing of non-verbal vocalisations of emotions is presented. The results show activation in pre-motor regions arising from passive listening to non-verbal emotional vocalisations, suggesting neural auditory-motor interactions in the perception of these sounds.

In sum, this thesis demonstrates that non-verbal vocalisations of emotions are reliably identifiable tokens of information that belong to discrete categories. These vocalisations are recognisable across vastly different cultures and thus seem to, like facial expressions of emotions, comprise human universals. Listeners rely mainly on pitch and pitch variation to identify emotions in non-verbal vocalisations, which differs with the cues used to comprehend speech. When listening to others' emotional vocalisations, a neural system of preparatory motor activation is engaged.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acknowledgements

Firstly I want to thank my supervisor Sophie Scott. She has supported and encouraged me in every possible way, as well as teaching me a million things on the journey: I truly could not imagine a better supervisor. Thanks also to Frank Eisner for his generous help, to Carolyn McGettigan and Jonas Obleser for their helpful feedback, and to the ICN Speech Communication Lab for being a lot of fun! I'm grateful to Andy Calder for useful feedback on earlier drafts, to Alastair McClelland for statistical advice, to Stuart Rosen for help with acoustics and stats, and to Mat Whalley for help with the fMRI analysis. My appreciation to Dacher Keltner and David Matsumoto for useful discussions, and Celia Heyes and David Carmel for always stimulating my mind. Thanks also to Richard Wise and the Hammersmith group, especially Jane Warren for help with the fMRI study. I'd also like to thank Debi Roberson and Jules Davidoff for their assistance in organising the Namibia project, and Marissa Nederhouser and Julie Goldstein for making it such a good trip. A special thanks to Paul Ekman for giving generously of his time and for his very valuable advice and feedback on my work. I would also like to thank the Department of Psychology at UCL for funding my PhD - and a special thanks to John Draper for being a star! Thanks also to Jade Wiland and Alex Kafkas for help with running experiments, and to all the participants who took part. And an additional thanks is due to Patrick Haggard and Elisabeth Hill for getting me into research in the first place. Also, thanks to friends and family for their support. Finally, and most importantly, I want to thank my partner, Alex Fradera. I really know no words to express my gratitude for his encouragement, love and help - Tack för allt.

1. BASIC EMOTIONS, DIMENSIONAL ACCOUNTS AND EMOTIONS IN THE VOICE.

This chapter begins by providing an overview of the debate on whether emotions are discrete or dimensional, an issue central to emotion research. I outline Ekman's theory of basic emotions and Russell's dimensional model, and discuss data from studies of facial expressions of emotions that support the two accounts. The discussion then turns to emotional communication in channels other than facial expressions, and then focuses specifically on the communication of emotion in human vocalisations. Different types of vocal expressions of emotions are delineated, with a particular emphasis on non-verbal signals. Data using vocal stimuli are discussed in the context of the theory of basic emotions and the dimensional model. I then outline research studying acoustic cues, culture, and neural processing of emotional vocalisations. Finally, the aims of this thesis are set out and discussed.

Emotions undoubtedly play a central role in all human lives, but what are they? Although everyone thinks they know what an emotion is, once asked, most find it difficult to give a definition (Fehr & Russell, 1984). In recent decades, psychologists have argued fiercely over what — if anything — emotions are. Theorists disagree on almost every aspect of emotions: what emotions are and do, whether they are adaptive or not, and how they relate to body, brain and culture. For example, the social constructionist school of thought argues that an emotion is merely a transitory social role, interpreted as a passion rather than an action (Averill, 1980). Some hold that emotions are biologically driven functions helping us to deal with our environment (Ekman, 2003). Others suggest that emotions are the result of a series of appraisals of pertinent information (e.g., Lazarus, 1991; Scherer, 2003). Indeed, it has been suggested that the concept of emotion should be eliminated altogether (Griffiths, 1997). No doubt some of these disagreements are the result of emotions being considered at many different levels of analysis: More than once have disputes between emotion theorists been compared to the fable of several blind men encountering different parts of an elephant and disagreeing on what the animal is truly like.

The disagreement on what emotions are has also influenced empirical research. Researchers with different views on emotions have tended to carry out rather different kinds of research. Whereas theorists with a relativist view have tended to define differences between (often quite similar) emotions in terms of contextual factors (e.g., Sabini & Silver, 1997), more biologically oriented researchers have tended to distinguish emotions on the basis of physiological variables (e.g., Levenson, Ekman, & Friesen, 1990) or facial muscle movements (e.g., Ekman, 2003). The theoretical framework in which emotion research is carried out thus affects the methodologies used. This relationship between theory and methodology is clear in the two most influential accounts of emotion, the theory of basic emotions and the dimensional model, which are outlined below.

1.1 Basic emotions

1.1.1 The foundations of the basic emotion account

According to the theory of basic emotions, emotions are evolved functions that change our physical and cognitive states in such a way as to help us deal with the cause of that emotion (Ekman, 1992a). Human beings are equipped with a set of emotions that are universal and innate, each of which evolved for their adaptive value in mobilising body and brain to deal with fundamental human tasks. Each basic emotion has a distinct physiological profile and a unique expression (Ekman, 1992b).

Darwin's book *The Expression of the Emotions in Man and Animals* (1872) provided the first account of emotions as evolved functions. He proposed that human emotional expressions had evolved to help us deal with our environment, but as our relationship to the environment had changed, these expressions had lost their original meaning. Darwin suggested that emotional expressions had acquired a secondary function in emotional communication between individuals, which had been retained as the primary function had disappeared. This was an important first step towards an evolutionary theory of emotions, but it would take over 80 years until Darwin's ideas were developed further. In the 1950s, Silvan Tomkins proposed that a small set of universal emotions were driven by innate subcortical programs linked to their expressions (Tomkins, 1955). Tomkins went on to formulate a two-factor theory of emotion: The biological factor formed the basis of a small set of universal emotions, while the culture-specific factor determined the social rules for displaying and inhibiting emotional expressions (Tomkins, 1962). Tomkins specified the face as the main focus of emotional expressions (Tomkins, 1962), which has remained central in the subsequent work on basic emotions.

1.1.2 A contemporary theory of basic emotions

The last 35 years have seen an explosion of interest in research on emotions, much of which has been heavily influenced by the theory of basic emotions. This is largely due to the work of Paul Ekman, the most important contemporary theorist of basic emotions. Ekman's view is largely consistent with the suggestions made by Darwin and Tomkins, although he has developed these ideas further and collected a wealth of empirical data to support the basic emotions account. He stated in his first paper on basic emotions that he agrees with Tomkins and with Darwin's idea "that there are distinctive movements of the facial muscles for each of a number of primary affect states, and that these are universal to mankind" (Ekman & Friesen, 1969). Ekman has also maintained the distinction that Tomkins made between biological and social factors of emotional functions. In contrast to Tomkins' proposal that the influence of cultural factors would be limited to social rules for displaying and inhibiting emotional expressions, Ekman hypothesised that the evokers of affect would also be culturally determined. According to Ekman's neuro-cultural model, culture determines not only the extent to which specific emotions are expressed, but also what triggers them (Ekman, Sorenson, & Friesen, 1969). Some triggers, such as snakes causing fear, would be universal, whereas others would vary between cultures. A number of other theorists have formulated alternative theories of basic emotions (e.g., Izard, 1977; Plutchik, 1980), but the fundamental concept of evolved, discrete, emotions, is present across these accounts.

1.1.3 Initial empirical evidence for the basic emotions

Ekman's account of basic emotions proposed that each of a number of affective states would be associated with a distinct set of facial movements, and the early studies of basic emotions consisted largely of asking naïve participants to match facial emotional expressions with verbal labels. Ekman carried out studies in several cultures and found high agreement in these samples (Ekman et al., 1969), whilst Carroll Izard independently collected similar data from eight cultures, using a different stimulus set (Izard, 1977). In order to combat criticisms that this agreement could be attributed to shared visual sources such as television and magazines, Ekman et al. also tested participants in visually isolated, preliterate cultures in Papua New Guinea and Borneo. Ekman et al. used photographs of Caucasian faces expressing the proposed basic emotions: anger, disgust, fear, happiness, sadness and surprise. The participants were shown each of the pictures in turn, selected by the experimenters to be especially strong expressions of the emotions, and were asked to pair the picture with one of five labels. The data from these preliterate cultures was similar to those from the literate cultures, although the effects were somewhat weaker. However, the methodology used was problematic in that it required the subjects to memorise the labels, and the visual isolation of the subjects was incomplete (Ekman & Friesen, 1971). In addition, the task relied heavily on correct translation of the sensitive emotion terms, which was problematic as it was difficult to ensure that the meaning of a given word was identical to the meaning of its nearest equivalent in another language.

To rectify these limitations, Ekman and Friesen tested a group of the Fore people in Papua New Guinea. These people were almost completely isolated from Western culture, and were unlikely to have been exposed to any images at all of Western emotional expressions. In order to create a task that did not rely extensively on working memory or verbal labels, the experimenters adapted a judgment task previously used with children (Dashiell, 1927). In this task, the participant is told a short emotion story and is asked to choose the one out of three photographs that best fits the story. The photographs were of emotional expressions which had previously been tested in literate cultures and been judged to be good exemplars, and the stories were specially formulated to be culturally appropriate for the Fore. The results were largely in line with the findings from literate cultures, providing strong support for the idea that certain emotions are universally associated with specific facial muscle actions (Ekman & Friesen, 1971). Ekman and Friesen then asked the Fore participants to make facial expressions to go with the six emotion stories. When the experimenters showed the Fore expressions to American participants, the Americans were able to consistently recognise expressions of anger, disgust, happiness and sadness (Ekman, 1972), providing further support for the basic emotions account. The exception to this pattern was a lack of discrimination between expressions of surprise and fear: The Fore participants did not differentiate between expressions of fear and surprise, and the American participants were unable to distinguish the Fore's expressions of fear and surprise. Ekman and Friesen hypothesised that in this case cultural factors may have affected the ability to make the surprise-fear distinction, in that the kinds of events that would be surprising in this cultural setting would be likely to also be perceived to be frightening. An alternative interpretation of this finding is that surprise is in fact not a basic emotion.

1.1.4 Later work on basic emotions

Ekman has continued to elaborate on a number of elements of the basic emotion theory; the meaning of "basic", the number of basic emotions and the exact characteristics that distinguish the basic emotions from other emotional phenomena and from one another (e.g., Ekman, 1992a; 1992b; 2003). A wealth of research has provided support for the basic emotion account. Cross-cultural work has confirmed that emotional expressions can be recognized across cultures (for reviews see Elfenbein & Ambady, 2002b; 2003), and a number of studies have detailed the facial muscle movements associated with each of the basic emotions (see Ekman & Rosenberg, 2005). Data from studies of developmental psychology (Denham, Zoller, & Couchoud, 1994; Smith & Walden, 1998), emotion in the elderly (Levenson, Carstensen, Friesen, & Ekman, 1991b; Calder et al., 2003) and psychopharmacology (Coupland, Singh, Sustrik, Ting, & Blair, 2003), have shown that expressions of one or more basic emotions can be selectively enhanced or impaired, implying that they comprise separate functional systems.

The majority of the literature on basic emotions has tended to focus on facial expressions of emotions, but a central feature of Ekman's basic emotion account is the suggestion that each basic emotion is associated with a specific pattern of autonomic arousal. Several studies have investigated this issue, using a number of different tasks, across several cultures (Ekman, Levenson, & Friesen, 1983; Levenson, Carstensen, Friesen, & Ekman, 1991; Levenson et al., 1990; Levenson, Ekman, Heider, & Friesen, 1992). These have employed a range of physiological measures, including heart rate, finger temperature, and respiratory depth. The consistent finding is that no single physiological measure differentiates between the basic emotions, but rather each basic emotion has a distinct physiological profile when all the physiological measures are considered together.

1.1.5 The negativity bias in the basic emotions

The basic emotions consist of a set of emotions rather strongly biased in terms of valence, with four negative (anger, disgust, fear and sadness), one neutral (surprise) and one positive emotion (happiness, sometimes called enjoyment or joy). Ekman has proposed that there may in fact be a number of positive emotions that constitute basic emotions, but that these would have distinct vocal expressions, while sharing the facial expression of the smile (Ekman, 1992b; 2003). The focus on facial expressions in basic emotion research has led to the positive emotions remaining largely unexplored (Ekman, 2003). One of the aims of this thesis is to test this hypothesis (see Section 1.10).

Whereas the basic emotion account holds that emotions are discrete functional systems, other theories view emotions as continuous entities that vary along central dimensions. Any attempt to discriminate between varieties of positive emotions needs to consider the possibility that they may not be distinct, but continuous.

1.2 The dimensional account

According to the dimensional view of emotions, affective states are not distinct, but are rather fuzzy concepts that relate to each other in a systematic fashion. These models have tended to focus on people's representations of emotions in terms of their internal affective space. The most influential contemporary dimensional account is James Russell's circumplex model (Russell, 1980), an elaboration of an earlier circumplex model by Schlosberg (Schlosberg, 1941; 1952). According to Russell's model, emotions are not distinct categories but rather vary along two independent, bipolar dimensions (Russell, 1980). The emotions are organised in a circle, a circumplex, which is placed on two lines representing the dimensions upon which they vary: valence and arousal (see Figure 1.1). According to the dimensional view, affective states are closely related to one another in a systematic way, which is in stark contrast to Ekman's perspective where the basic emotions are essentially mutually exclusive categories engaging different, independent processes. This disagreement will be discussed in more detail later.

In Russell's model the emphasis is on continua: emotions have fuzzy edges and are related to other emotions in proportion to their distance in emotional space. The valence dimension refers to the pleasantness of the subjective emotional state, with delighted, happy, pleased and glad at the top end, and sad, miserable, frustrated and distressed at the bottom. Arousal refers to a sense of energy of the emotion, with alarmed, tense, aroused and astonished at one extreme, and calm, sleepy, tired and droopy at the other (Russell, 1980). Russell acknowledges that the model includes many states that are not actual emotions (e.g., fatigue, sleepiness and placidity), but maintains that it nevertheless provides a description of affective experience, which always consists of a blend of pleasure and arousal (Russell & Feldman-Barrett, 1999). A number of other theorists have presented dimensional models more or less similar to Russell's (e.g., Davidson, 2000; Watson &

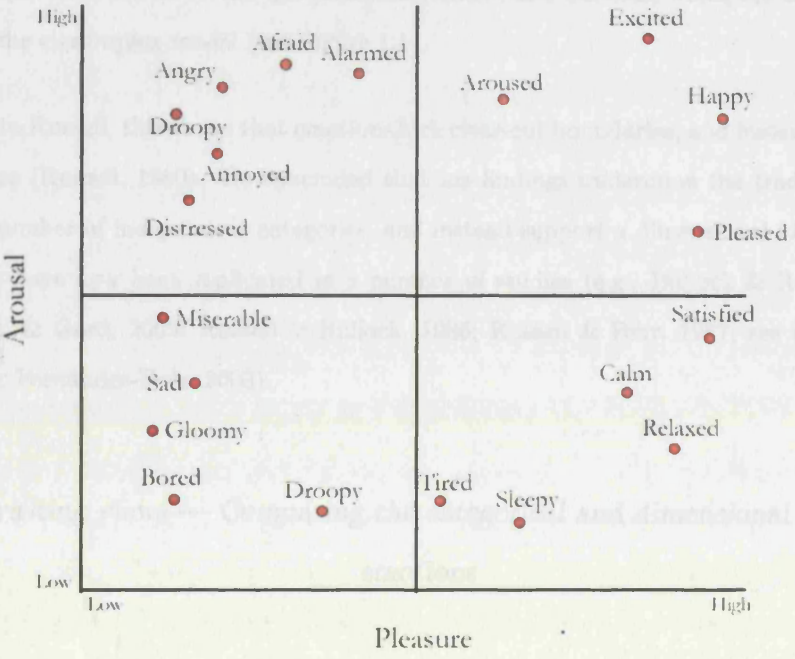

Fig. 1.1: Russell's circumplex model of affect. Adapted from Russell, 1980.

Tellegen, 1985), but as the circumplex model has been the most influential of these in contemporary emotion research, the current discussion of dimensional theories will be limited to this account.

1.2.1 Empirical support for the dimensional account

Much of the work supporting the dimensional account of emotions has been aiming to create a semantic map of affective space using emotion words. In his initial study, Russell selected a set of 28 words to represent the realm of affect, which participants were asked to sort into eight categories. According to Russell's hypothesis, the eight categories represent various points on the dimensions of arousal and valence, evenly distributed in the circumplex model. Consistent with his model, Russell found that the emotion words were sorted into the predicted categories. The participants were then asked to place the category labels into a circular order, based on proximity of the meaning of the terms. Russell found that the subjects ordered the category labels in a way that was consistent with dimensions of valence and arousal. He also noted that across subject, each word was placed into more than one category, and that these "errors" varied systematically with the distance in the model from the most commonly used category. For example, the word tense was most often placed in the Arousal category, but on the instances where it was not placed there it was most often placed in the categories Excitement and Distress, which are on either side of Arousal in the circumplex model (see Figure 1.1).

According to Russell, this shows that emotions lack clear-cut boundaries, and instead constitute fuzzy categories (Russell, 1980). He concluded that his findings undermine the traditional view of affect as a number of independent categories, and instead support a dimensional view of affect. These findings have now been replicated in a number of studies (e.g., Bullock & Russell, 1985; Kring, Barrett, & Gard, 2003; Russell & Bullock, 1986; Russell & Fehr, 1987; see also Russell, Bachorowski & Fernández-Dols, 2003).

1.3 Contrasting views — Comparing the categorical and dimensional views of emotions

1.3.1 Categorical perception

The basic emotion view emphasises the biological elements of a small set of emotions and their evolutionary origins, and focuses on the communication of emotions with the use of emotional expressions. The dimensional accounts of emotions focus on people's representations of emotions, arguing that emotions are what people construe them to be. Although these ideas differ profoundly, they need not be mutually exclusive, and could be considered descriptions at different levels of analysis. Nevertheless, proponents of both theories argue fiercely with each other, with the main point of disagreement being whether emotions comprise a set of distinct systems or whether they are variations within a single system. Researchers from the two schools of thought have carried out quite different types of research: the Ekman school has focused primarily on categorisation studies using facial stimuli, whereas Russell and his colleagues have used emotion words with participants grouping them or rating them for similarity. However, a number of studies have directly compared the two accounts using the categorical perception paradigm (Calder, Young, Etcoff, Perrett, & Rowland, 1996; Etcoff & Magee, 1992; Young, Rowland, Calder, Etcoff, Seth & Perrett, 1997). Categorical perception means that a change along a continuum is perceived as an abrupt change between distinct categories, rather than as a gradual shift.

Etcoff and Magee (1992) used computer-generated line drawings of faces to study categorical perception of emotional expressions. They found that when participants were asked to discriminate pairs of faces, faces within a category were discriminated more poorly than faces in different categories, even though the physical difference between the two stimuli was held constant. They interpreted this to mean that "emotional expressions, like colors and speech sounds, are perceived *categorically*, not as a direct reflection of their continuous physical properties" (p. 227; italics in original). Calder, Young, Etcoff et al. (1996) replicated and extended Etcoff and Magee's findings, using photographs of emotional facial expressions rather than line drawings. Calder et al. found that discrimination was better for pairs of stimuli that belonged to different categories than for two stimuli belonging to the same category, providing additional support for the categorical perception of facial expressions of emotions.

In a more extensive study using morphed photographic images of all pair-wise combinations of the six basic emotions and neutral expressions, Young et al.(1997) replicated the findings by Calder, Young, Etcoff et al., and showed that this pattern was also found for neutral expressions. Young et al. drew attention to a number of ways in which their findings are inconsistent with dimensional accounts of emotions. For example, according to Russell's model, an expression that is half way between happiness and anger should be perceived as expressing surprise. This was not supported by Young et al's findings. Rather, the expression was perceived as expressing happiness up a certain point, after which was perceived as expressing anger. Young et al. also point out that, according to a dimensional account, participants should have great problems categorising expressions that are morphs of two emotions that are on opposite ends in emotional space. Dimensional accounts would predict that these expressions would be perceived as neutral, but this has was not found to be the case, even when the label "neutral" was available (Young et al., 1997, Experiment 2).

Categorical perception of emotional expressions has now been demonstrated further with 7-month old infants (Kotsoni, de Haan, & Johnson, 2001), using event-related potentials (Campanella, Quinet, Bruyer, Crommelinck & Guerit, 2002) and with emotional speech (De Gelder & Vroomen, 1996; Laukka, 2005; see Section 1.5.1 for a discussion of categorical perception in emotional speech). These findings lend support to the basic emotion view that boundaries between emotions are sharp rather than fuzzy. However, other research assessing the basic emotion account and dimensional models have found support for the latter.

1.3.2 The role of context

According to the theory of basic emotions, observers interpret emotional expressions directly in terms of specific emotional categories (Tomkins, 1962). In contrast, Russell's account proposes that when one is viewing someone else's emotional expression, one does not infer the other person's specific emotion. Instead the viewer perceives the sender's state in terms of arousal and valence, as well as receiving quasi-physical information, such as smiling or weeping (Russell, Bachorowski & Fernández-Dols, 2003). Specific emotions are attributed when this inference is combined with contextual information.

Carroll & Russell (1996) designed a study to contrast the two perspectives in terms of the role they attribute to situational information when interpreting emotional expressions. The basic emotion position holds that the interpretation of others' facial expressions should be a direct function of the expression itself, regardless of contextual information. In contrast, Russell and Carroll proposed the "limited situational dominance" position. They proposed that there would be cases where viewers would base their emotional judgments on situational cues. Specifically, this would be the case if the facial expression and situational cues were incongruent, but similar in terms of valence and arousal.

A number of previous studies had failed to find support for any situational influence on facial expression judgements (e.g., Nakamura, Buck & Kenny, 1990; Watson, 1972). Carroll and Russell examined the established paradigm, in which a set of faces would be paired with each and every one of a set of stories. They hypothesised that the way in which the facial stimuli and stories were matched could affect the result of the study. In Carroll & Russell's study, stories were designed to match the facial expressions they were to be paired with on quasi-physical information and pleasantness and arousal, but to differ in the specific emotional state. For example, a fearful face would be paired with a story that would describe an anger situation, but never a happy situation. Carroll and Russell expected people to respond "anger" on the basis that the facial expression would primarily be interpreted as expressing a highly aroused, negative state in which the person is staring (quasi-physical information).

The results of the study showed that participants based their judgements on the emotion expressed in the story rather than the face shown, confirming their hypothesis and Russell's dimensional model of emotions (Carroll & Russell, 1996). Unfortunately, the study only included comparisons of emotions that were close to each other in the circumplex model. In light of Carroll and Russell's argument that their findings were due to a circumplex-based mental representation of emotion, it would have been useful to also include comparisons where their hypothesis would have predicted that participants would base their judgments on the facial expressions: if participants also based their judgments on situational cues where situational and facial expression cues differed greatly in terms of arousal and valence, this would not support Carroll and Russell's conclusions. Including a wider range of face-story pairings would have allowed them to test whether the basis of participants' choices actually corresponded to variations on the dimensions that they used to explain their findings. Nevertheless, Carroll and Russell's findings do lend some support for the influence of situational factors on the interpretations of emotional facial expressions.

1.3.3 Contrasting views: Summary

Much of the research investigating emotional processing has tended to be heavily influenced by one or other of the dominant theoretical perspectives, with little cross-talk between proponents of the different views. Few studies have tried to compare predictions from the different accounts, although the growing use of the categorical perception paradigm is promising. According to the late emotion theorist Richard Lazarus, "such disputes are never completely settled by the data, but are resolved over time by how generative each position is — what new questions and new findings each has led to" (cited in Ekman, 1994, p. 283). The current thesis cannot attempt to prove or disprove either account, but the research should be considered in this framework of emotion theories.

1.4 Different channels for communicating emotions

The majority of work on emotional communication has focused on facial expressions. However, human beings have a rich repertoire of other signals with which to communicate their emotions. Before the more detailed review of research of work investigating emotional vocalisations, a brief discussion follows on signals other than facial expressions that are used to communicate emotions.

1.4.1 Communicating emotions with body language

Body language is an efficient means of communicating both an emotional state and an intended action concurrently (de Gelder, 2006). High recognition accuracy can be obtained from brief clips of body motion communicating fear, anger, grief, joy, surprise, and disgust (Dittrich, Troscianko, Lea & Morgan, 1996). Recognition is best from fully lit scenes but remains above chance even with point-light displays, where the viewer only sees a number of small points of light attached to the body of the sender. Dittrich et al. also found that exaggerated body movement improved recognition accuracy and was rated as more emotionally intense, suggesting that body movement influences not only what emotion is perceived but also to what degree. However, movement is not necessary for emotion to be perceived from body displays: A study by Atkinson, Dittrich, Gemmell, and Young (2004) replicated Dittrich et al.'s finding and also showed that emotions can be recognized from static displays, although recognition accuracy is not as high as for dynamic stimuli.

1.4.2 The smell and taste of emotions

Little work has studied the perception of emotions in the olfactory and gustatory modalities. An exception is a study by Vernet-Maury, Alaoui-Ismali, Dittmar, Delhomme and Chanel (1999), which showed that basic emotions can be induced from smells. They asked participants to inhale a range of pleasant, neutral and unpleasant odorants, while recording their autonomic nervous system activity. The patterns of autonomic responses obtained for each subject and each odorant were transcribed into one of the six basic emotions using a decision tree built from previous autonomic analyses (e.g., Ekman et al., 1983). They found a distinct set of autonomic responses corresponding to each of the basic emotions, and concluded that "autonomic changes can identify the quality of the response, i.e. the type of basic emotion" (p. 182). A recent study by Robin, Rousmans, Dittmar & Vernet-Maury (2003) used a similar methodology to study the induction of emotions from tastes. Participants tasted sweet, salty, bitter, and sour solutions while their autonomic nervous activity was monitored. Similarly to the emotions induced by smells, the researchers claim that taste can elicit patterns of autonomic responses that correspond with those previously found for each of the six basic emotions. The researchers pointed out that there was a substantial amount of variation in the emotions induced in each subject for each taste, but that the relationship between the emotion and the autonomic activity was consistent.

1.4.3 Vocal expressions of emotions

> "With many kinds of animals, man included, the vocal organs are efficient in the highest degree as a means of expression" (Darwin, 1872/1998, page 88).

The human voice is a highly complex tool of communication: We use our voice not only to transmit linguistic meaning, but also to communicate to others our feelings, sex, age, body size, social class, and geographical origin (Karpf, 2006). One single type of information, such as emotions, can be communicated using a number of cues: Emotions can be conveyed using semantic cues, affective prosody, or non-verbal vocalisations such as laughter and sighs. In this thesis, the discussion will focus mainly on the non-verbal aspects of emotional vocalisations.

The voice can be used to communicate emotions in segments of different lengths. A non-verbal signal could consist of a brief sigh or a lengthy laugh. In a short verbal exclamation such as "oh" or "wow", the affective value of the expression is contained within the acoustic cues of a single syllable. In contrast, in a sentence, the prosodic melody throughout the sentence can be used, including such features as the speed of speaking and the stress placed on individual words. Murray, Arnott and Rohwer (1996) point out that emotional speech at the sentence level tends to vary in terms of three important parameters: voice quality, pitch contour and timing. Voice quality refers to the "character" or timbre of the voice, for example whether it sounds whispered or creaky. Pitch contour refers to the intonation of an utterance, especially the range of pitch. The timing includes the overall speed, as well as changes in duration of certain parts of the utterance. Some work on emotional vocalisations has been done using sentences (e.g., Kucharska-Pietura, David, Masiak & Phillips, 2005; Laukka, 2004; Mitchell, Elliot, Barry, Cruttenden & Woodruff, 2003; Wildgruber, Pihan, Ackermann, Erb & Grodd, 2002), while other studies have used single words (e.g., Buchanan et al., 2000). Relatively long stimuli such as sentences can be advantageous for study, as the stimuli include a fuller range of acoustic cues, and the stimuli arguably have high ecologic validity. On the other hand, longer stimuli contain more variation and thus more noise in the signal. In addition, single words allow easier matching for acoustic features which may not be of interest, such as duration.

One disadvantage with using speech as stimuli is that it contains semantic information. This may be congruent or incongruent with the paralinguistic content of the signal, and hence interfere or facilitate the listener's judgment of the emotional tone of the speech. This is a problem particularly in studies using real speech samples (Laukka, 2004). In addition, the content of the verbal utterance can affect the acoustic parameters of the sound which carries the emotional information (Banse & Scherer, 1996). To avoid these confounds, researchers using speech stimuli have tended to use the "standard contents paradigm" (Davitz, 1964), where the same sentence or word is produced several times, each time expressing different or no emotions.

Another way to ensure that semantic information does not confound the stimuli is to use pseudo-speech (e.g., Banse & Scherer, 1996; Grandjean et al., 2005; Sander et al., 2005). For example, Banse and Scherer used two sentences composed of phonemes from six Western European languages: "Hat sundig pron you venzy" and "Fee gott laish jonkill gosterr". According to the authors, the sentences are perceived as speech in a foreign language. This suggests that similarly to real speech, pseudo-speech is processed within the framework of phonetic processing, making it a useful way to tap into the paralinguistic signals used in speech without interference due to semantic information.

Other researchers have used entirely non-verbal vocalisations of emotions, such as screams, sighs and laughs (e.g., Phillips et al., 1998; Scott, Young, Calder, Hellawell, Aggleton & Johnson, 1997). These kinds of stimuli would be unlikely to be processed like speech, as they contain minimal phonemic information, and thus arguably form better analogues to other non-verbal emotional stimuli such as faces. One advantage of non-verbal emotional vocalisations is that they are perceived to be more spontaneous and reliable signals of emotion (Johnstone & Scherer, 2000). Their ecological validity is relatively high, as these are sounds that are used in everyday life. In contrast, stimuli consisting of emotions overlaid on speech or pseudo-speech are an artificial merge of two elements, emotions and phonemes. Of course, speech is infused with affect in everyday life, but not within a standard contents paradigm. One study has used vocal stimuli with a range of phonemic information (Schröder, 2003), referred to as 'affect bursts' (Scherer, 1994). Affect bursts denotes a range of emotional vocalisations, including 'raw affect bursts' such as laughter and 'affect emblems' such as "yuck". The study by Schröder is discussed in more detail in Chapter 2.

With the exception of one chapter (Chapter 5), the focus of this thesis is on the processing of non-verbal vocal expressions of emotions containing a minimum of phonemic information. These will be referred to as "non-verbal vocalisations of emotions" or "non-verbal expressions of emotion" to distinguish them from the broader class of 'affect bursts' (Scherer, 1994). Given the dearth of research into non-verbal vocal expressions of emotion, most of the discussion of work on emotional vocalisations will focus on studies using emotional speech or pseudo-speech stimuli, with research using non-verbal vocalisations included where available. Following on from the previous discussion on categorical and dimensional accounts of emotions, the next section discusses a number of studies that have used vocal expressions of emotions to support either account or to differentiate between them.

1.5 Categorical and dimensional accounts of emotional vocalisations

1.5.1 Vocalisations of basic emotions

Tomkins (1962) and later Ekman (1992b) have postulated distinct vocal signals for each of the emotions which have distinctive facial expressions, that is, anger, disgust, fear, happiness, sadness, and surprise. As already mentioned, the vast majority of research into the basic emotions has used facial stimuli and thus the investigations attempting to empirically test this hypothesis have been sparse. The focus of this work has tended to be on the processing of sounds of fear, and to some extent, disgust.

A study by Scott et al. (1997) tested emotion recognition in a patient with bilateral amygdala lesions and controls, using verbal and non-verbal vocalisations of the basic emotions. They found that both verbal and non-verbal vocalisations of the basic emotions could be reliably recognized by the control participants, supporting Ekman's hypothesis that there are distinct vocal signals for each of the basic emotions. Scott et al. also showed that the patient with amygdala damage was specifically impaired on recognition of vocalisations of fear and anger, which has previously been demonstrated in similar patients using visual stimuli (Adolphs, Tranel, Damasio and Damasio &, 1994). These selective impairments have been interpreted as supporting the notion of a set of emotions that constitute separate functional processes, as suggested by the basic emotions account.

Several imaging studies have shown that the perception of vocal expressions of different basic emotions involves different neural structures. A study by Morris, Scott, and Dolan (1999) used Positron Emission Tomography (PET) to examine listeners' neural activity during exposure to non-verbal vocal expressions of basic emotions. The study found that processing of auditory stimuli expressing fear involved similar brain regions that have previously been implicated in visually stimulated fear processing, most notably the amygdala. A study by Phillips et al. (1998) used functional Magnetic Resonance Imaging (fMRI) with six male normal participants to investigate the neural processing of facial and non-verbal vocal expressions of fear, disgust and neutral expressions. They found that both kinds of fearful stimuli activated the amygdala, whereas facial, but not vocal expressions of disgust activated the anterior insula and the caudate-putamen. These results from these studies indicate that the amygdala is involved in processing fearful stimuli from both visual and auditory input, whereas the areas associated with processing disgust signals may not be modality independent. However, more research into the neural processing of vocal expressions of emotions is needed (see Chapter 6 for a more extensive discussion on functional imaging work on emotion communication).

There is some support for vocal expressions of the basic emotions from research into the acoustics of emotional vocalisations. In a review of research on vocalisations of the basic emotions, Murray and Arnott (1993) pointed out that the findings are largely consistent between authors and studies, indicating that vocalisations of each the basic emotions is associated with a distinct pattern of acoustic cues, just as the facial expressions of the basic emotions are associated with particular muscle movements (Ekman & Rosenberg, 2005). Murray and Arnott conclude that the available data supports the notion of consistent, specific vocal profiles for each of the basic emotions, although they point out that more studies are needed.

Some research has investigated the universality of vocal signals of emotions. The most extensive study to date was carried out by Scherer, Banse, and Wallbott (2001), who tested recognition of vocal expressions of anger, fear, sadness, joy and neutrality. They used sequences of European language nonsense-syllables (see section 1.4.3) with participants in nine countries in Europe, North America and Asia. They found an overall recognition rate of 66% and strong inter-cultural consistency in the types of errors participants made. This study is discussed in more detail in Chapter 4, but Scherer et al.'s findings demonstrate cross-cultural consistency in emotion recognition from vocal stimuli, thus lending support to the basic emotion account.

Additional evidence comes from a study by Laukka (2003), which demonstrates categorical perception of emotions in speech. Laukka used concatenative speech synthesis, incorporating acoustic elements from emotional speech into a neutral sentence to create six continua (anger-fear, anger-sadness, fear-happiness, fear-sadness, happiness-anger, and happiness-sadness) of emotional speech. He used an ABX task, in which the participant has to select which one of two stimuli, A or B, best matches a third stimulus, X. Laukka found that "each emotion continuum was perceived as two distinct sections separated by a category boundary" (p. 284), evidence for categorical perception. The participants also performed a forced choice task, which showed that listeners identified stimuli with the prototypes at either end of the relevant continuum, with a sudden shift in judgments in the middle of the continuum. These data demonstrated that perception of vocal expressions of emotions is categorical. Laukka suggests that categorical perception of vocal emotion expressions may have evolved in order to facilitate the rapid classification of others' states, a suggestion in line with the basic emotion account.

1.5.2 Dimensions in vocal expressions of emotions

Vocal expressions of emotions have often been assumed to communicate only physiological arousal rather than specific emotions (see Scherer, 1986). Russell et al. (2003) state that the receiver of a vocal emotional signal perceives the state of the sender in terms of the bipolar dimensions of arousal and to a lesser degree valence, rather than as categorical emotion states.

Some empirical work, mainly focused on speech production, has supported this view. Bachorowski (1999) claims that "the most parsimonious interpretation of production-related data is that speech acoustics provide an external cue to the level of non-specific arousal associated with emotional processes." (p. 55). She cites a study by Bachorowski and Owren (1995), where participants were given fake positive or negative performance feedback during the performance of a difficult lexical decision task. After each feedback presentation the subjects were asked to pronounce the words "test n test", where n was the number of the next trial. The variations in F_0, jitter (irregularities between successive glottal pulses perceived as pitch perturbations) and shimmer (small variations in amplitude maxima perceived as loudness perturbations) were analysed. Bachorowski and Owren found that subjects' vocal expressions were affected by the feedback they received and that this interacted with their emotional intensity ratings. A follow-up study using a similar methodology further established that participants' self-reported arousal affected the participants' vocalisations (Bachorowski & Braaten, 1994). It is worth noting that neither the study by Bachorowski and Owren nor the one by Bachorowski and Braaten actually investigated whether the acoustic cues would map onto *specific* emotions, but only investigated their relation to dimensional properties. A recent study by Bänziger and Scherer (2005) provides supporting evidence for the dimensional account. Bänziger and Scherer investigated the role of F_0 (pitch) contour in the perception of emotional vocalisations in nonsense-speech. The F_0 level was affected by the level of emotional arousal, but not by the specific emotion of the sound. It is notable that the studies that have provided support for the view that vocalisations communicate information about arousal rather than specific emotional states have tended to examine a very limited set of acoustic cues.

1.5.3 Summary: Categorical and dimensional views of emotional vocalisations

The research into emotional vocalisations has lent support to Ekman's (1992b) proposal that there are vocal correlates of the basic emotions. Research shows that emotional vocalisations can be reliably recognized across cultures (Scherer, Banse & Walbott, 2001; see also Chapter 4 for a more extensive discussion of this issue). Neuropsychological data has found that brain damaged patients whose recognition of certain basic emotions in the visual domain are selectively impaired show similar impairment in the recognition of vocal expressions of those same basic emotions (Scott et al., 1997). This is complemented by functional imaging data showing selective activation during perception of different basic emotions (Morris et al., 1999; Phillips et al., 1998). These findings are promising for the basic emotion account, although research into emotions others than fear and disgust is somewhat lacking. Finally, a study by Laukka (2003) has shown evidence for categorical perception of emotional vocalisations of the basic emotions.

According to Bachorowski (1999), research from her group has shown that emotional vocalisations mainly communicate emotional arousal, providing some support for Russell's dimensional account. One problem with this interpretation is that these studies did not directly assess the extent to which specific emotional states could be inferred from the vocalisations. A more recent study by Bänziger & Scherer (2005) did examine this issue and found that the F_0 level of emotional sounds was influenced by the speaker's state of emotional arousal but not by specific emotions. Together, this data seems to support Russell's dimensional model, although more studies explicitly comparing the two accounts are needed. Laukka (2004) has suggested that emotions may be categorical but that emotional dimensions may correspond to intellectual (but not perceptual) emotion constructs. This is an attractive idea, albeit difficult to reconcile with the data showing that emotional arousal is reflected in acoustic voice cues.

A number of methodologies have been used to investigate emotional vocalisations, and this thesis employs several of these. The following section provides an outline of work into the roles of acoustics and culture in vocalisations of emotions, and the neural processing of these kinds of signals.

1.6 The acoustics of emotional vocalisations

Research in the domain of facial expressions of emotions have established not only that participants can infer emotions from facial expressions, but also the specific muscle movements used to signal different emotions (Ekman & Rosenberg, 2005). Although a number of studies from different laboratories have now shown that naïve listeners can identify emotions from emotional vocalisations (Juslin & Laukka, 2003; Scherer, 2003; see Chapters 2 and 5), little is known about which cues are used in vocal communication of emotions (Juslin & Scherer, 2005). This is likely due in part to the difficulty in measuring many of the acoustic cues that are thought important (Scherer, 1986). Nevertheless, several studies have investigated the acoustic cues in emotional speech or pseudo-speech (e.g., Banse & Scherer, 1996, Bänziger & Scherer, 2005; Laukka, 2004; see Murray & Arnott, 1993), although most studies have only included a small set of acoustic features (e.g., Bänziger & Scherer, 2005; see Juslin & Laukka, 2003). This work is discussed in more detail in Chapters 2 and 5, but an overview is given here.

Pitch is thought to be a key feature for the communication of emotion in speech. In an early review of emotion in speech, Murray and Arnott (1993) found that the pitch envelope was the most important parameter for differentiating between the basic emotions. In contrast, a recent study by Bänziger & Scherer (2005) found that F_0 levels of emotions sounds were affected by the level of emotional arousal, but not the specific emotion of the sound.

Two studies have investigated whether participants' perception of emotional sounds can be predicted from the acoustic cues of the sounds. Banse & Scherer (1996) measured a number of acoustic cues of emotional nonsense-speech. They regressed the acoustic cues onto participants' use of emotion categories for each stimulus class in a forced-choice paradigm. They found that for most of the emotions, the participants' classification could be predicted by some constellation of the acoustic cues. Laukka (2004) measured a set of voice cues from speech expressing five emotions. He performed a series of multiple regressions with nine acoustic measurements including F_0, speech rate, and mean voice intensity as independent variables, and the participants' ratings of the sounds on emotional rating scales as dependent variables. All of the ratings could be significantly predicted by the acoustic cues of the sounds, with a unique constellation of cues predicting each emotion.

1.6.1 Scherer's componential theory of acoustic cues in emotional speech

Scherer has proposed a componential model of emotion, which makes predictions about the acoustic patterns of emotional speech (Scherer, 1986; 2001; 2003). According to Scherer's theory, emotions are the result of a series of stimulus evaluation checks. These affect the nervous system, which in turn influences aspects of the vocal production system. On the basis of this model, Scherer has made a set of predictions of the acoustic properties expected for vocal expressions of a number of emotions. The current thesis will not attempt to test the predictions of Scherer's model. Many of the acoustic parameters specified in Scherer's model are difficult to measure empirically (e.g., F2 mean, formant 'precision', spectral noise), and not well suited for use with the kinds of nonverbal stimuli used in most of this thesis. A systematic empirical test of Scherer's predictions would constitute a substantial and challenging feat, which is likely the reason it has not yet been attempted.

1.7 Cross-cultural work on emotional vocalisations of emotions

Little work has studied communication of emotions cross-culturally. The only extensive study to date was carried out by Scherer, Banse, & Wallbott (2001). They tested recognition of pseudo-speech expressing anger, fear, sadness, and joy in students from seven European countries, the USA and Indonesia. The stimuli were sentences made up from combinations of syllables from six European languages. Scherer et al. found an overall recognition rate of 66%, demonstrating that emotional expressions can be recognized from vocal stimuli. However, there was substantial variation in the recognition rates found across the different cultures. Notably, the participants from Indonesia performed worst on the task. One problems in interpreting these results is that the Indonesian participants were not only culturally less similar to the other groups than the other groups were to one another, but they were also the only group whose native language was not related to any of the languages used to produce the stimuli. Thus, the phonemic contents of the stimuli may have biased the results in favour of the participants who spoke languages related to the languages used to produce the stimuli. This finding shows the limitations of using speech-like stimuli based on phonemes from only a handful of languages, as differences in culture often co-occur with differences in language.

However, research studying facial expressions of emotions has also found that the highest recognition scores tend to be found in cases where the participant and the poser (stimulus producer) are from the same culture, with recognition gradually decreasing with an increase in cultural distance between poser and participant (Elfenbein and Ambady, 2002b). In order to explain this in-group bias, Elfenbein and Ambady (2003) have proposed a dialect account of emotional communication. According to this model, emotional expressions are universal, but each cultural group has some minor culture-specific variation on the original expressions. These specific adjustments are acquired by social learning. With increasing cultural distance the participant would have less exposure to the variations employed by the poser and thus be less good at recognizing them (Elfenbein and Ambady, 2003). This issue will be discussed in more detail in Chapter 4.

1.8 The neural processing of emotional vocalisations

As in the broader field of emotional communication, the cognitive neuroscience of emotions has tended to focus mainly on the processing of facial expressions of emotions, and insofar as emotion in the voice has been studied, this work has mainly been done using emotional speech. Much of the debate has focused on the extent to which processing of emotional prosody in speech is right lateralised (see Pell, 2006 for a discussion). Several accounts have been proposed that argue for a hemispheric lateralisation in the processing of emotionally inflected speech, some focused on temporal aspects of speech and others on affective components. Some authors have suggested that the right temporal region is involved in the analysis of slow acoustic variations, for example supra-segmental information in speech, such as prosody, whereas the left temporal region mainly processes rapid acoustic changes, such as phonemes (Poeppel, Guillemin, Thompson, Fritz, Bavelier & Braun, 2004; Zatorre, 2001). Davidson (1998) has instead suggested that lateralisation in the processing of emotional stimuli would be based on the valence of the stimuli, such that positive stimuli are processed mainly in the left hemisphere and negative stimuli are processed mainly in the right hemisphere. This is based on a dimensional model of emotions proposed by Davidson (1992), hypothesising that emotions are a function mainly of approach- and withdrawal-related behaviour.

Although a more extensive review of existing work on the neural processing of emotional vocalisations is given in Chapter 6, a brief review of the neuropsychological and functional imaging data is given here. A number of studies using brain damaged patients have found evidence that emotional speech depends primarily on right hemisphere structures (e.g., Adolphs, Damasio & Tranel, 2002; Blonder, Bowers & Heilman, 1991). However, a recent study by Pell (2006) examined emotion recognition from prosody in an extensive group of patients with focal lesions to either the right or left hemisphere. He found that both types of patients were impaired in their comprehension of prosody, but that right hemisphere damage was associated with an insensitivity to emotive features in the speech, whereas left hemisphere patients had difficulty interpreting the prosodic cues in a language context.

A number of studies have used functional imaging techniques to investigate the neural processing of emotional vocalisations. Most of these have tended to compare the processing of neutral with emotionally inflected speech or pseudo-speech (e.g., Grandjean et al., 2005; Mitchell et al., 2003; Wildgruber et al., 2002). Notably, these studies have not always controlled for differences between stimulus categories in low-level acoustic features, such as pitch variation. Findings have tended to show either a right-lateralised activation (e.g., Mitchell et al.) or bilateral activation, with somewhat more activation on the right than the left (e.g. George, et al., 1996). There is some variation in the specific regions that have been found to be involved in the processing of emotional prosody, but activation has primarily involved temporal regions including the superior temporal sulci (Grandjean et al., 2005; Meyer, Zysset, von Cramon & Alter, 2005) and temporal poles (Imaizumi et al., 1997; Mitchell et al., 2003). Activation has also commonly been found in frontal areas including the orbitofrontal cortex (Sander et al., 2005; Wildgruber, Hertrich, Riecker, Erb, Anders, Grodd & Ackermann, 2004). Some subcortical activation including the amygdalae (Sander & Scheich, 2001) and caudate (Kotz, Meyer, Alter, Besson, von Cramon & Friederici, 2003) has also been reported. Pell (2006) argues that although the available evidence suggests that the right hemisphere is especially important in the processing of emotional prosody, it seems that the interpretation of emotion in speech recruits an extensive network including structures in both hemispheres (see also Schirmer & Kotz, 2006).

1.9 The focus on negative emotions

Generally, theory and research on the psychology and neuroscience of emotion have been oriented around negative affect (Berridge, 2003; Fredrickson, 1998). It has been suggested that "psychologists have inadvertently marginalized the emotions, such as joy, interest, contentment, and love, that share a pleasant subjective feel" (Fredrickson, 1998, p. 300). Although a few specific areas of positive affect (e.g., laughter and sexual pleasure) have received some attention from affective neuroscience (see Fried, Wilson, MacDonald, Behnke, 1998; Karama et al., 2002; Rodden, Wild, Erb, Titze, Ruch & Grodd, 2001), a more systematic investigation into positive affect is lacking. Specifically within vocal expressions of emotion, there have been calls for distinctions to be made between different positive states: "As has become painfully clear in the attempt to review the vocal expression literature above, a comparison of results from different studies is virtually impossible if it is unclear whether ... 'happiness' refers to quiet bliss or bubbling elation" (Scherer, 1986, p.163).

As mentioned earlier, Ekman has put forward the hypothesis that there may be several positive emotions that share the facial expression of the smile, but that each have unique vocal expressions (Ekman, 1992b; 2003). For facial expressions of emotions, it is typical to find higher recognition rates for judgments of happy expressions than other emotions (Ekman, 1994; Elfenbein & Ambady, 2002b). In contrast, participants typically have difficulty identifying vocal expressions of joy/happiness from vocal stimuli, a pattern found across a number of cultures (Scherer, Banse, & Wallbott, 2001). One interpretation of this pattern is that there may be several positive emotions that share the facial expression of the smile, but possess distinctive vocal expressions (Ekman, 1992b; 2003). The need for distinction within positive emotion categories is further supported by findings from a study of emotional vocalisations which included one more intense and one less intense form of seven emotions (Banse & Scherer, 1996). It was found that unlike all the other within-emotion pairs (e.g., sadness-despair, anxiety-panic) the two positive emotions (elation and happiness) were virtually never confused with each other, suggesting that what were called elation and happiness may be two distinct emotions rather than two strands of the same emotion, happiness.

1.10 Aims of this thesis

In the preceding discussion I have outlined research investigating a number of aspects of emotional vocalisations. I have identified several issues that have not been addressed in the research done to date, and these issues form the basis for this thesis. I have pointed out the lack of research into positive emotions, and suggested that the reason for this may be the focus emotional communication research has had on facial expressions. This focus has been at the expense of vocal signals, which may be where these distinct positive emotions are most clearly communicated (Ekman, 1992b). I have also pointed out that there is a particular dearth of research into non-verbal emotional expression, that is, vocalisations that do not contain speech or phonemes, even though these kinds of signals may be highly reliable indicators of emotion (Johnstone & Scherer, 2000). The nature of emotions as categorical or dimensional entities is currently unresolved, and the domain of non-verbal emotions is a relatively unexplored arena in which to further this debate. Furthermore, the issue of how acoustic information affects the perception of emotion in both speech and non-verbal stimuli is poorly understood, as is the relationship between speech comprehension and emotion recognition in speech. I have described how the influence of culture on emotion processing from vocal stimuli has been almost entirely neglected, but would serve to shed light on debates as to the universality of non-verbal vocal signals of emotion. Finally, we need a systematic investigation into the neural correlates of the perception of non-verbal vocalisations. In the next section I will set out how this thesis will attempt to address these issues. The main aims of this thesis are structured as a set of questions which this thesis sets out to investigate.

Can vocalisations of positive emotions be reliably identified?

One of the aims of this thesis is to empirically test the hypothesis put forward by Ekman (1992b) that there is a set of positive emotions with distinct, recognisable expressions. The specific set investigated are achievement/triumph, amusement, contentment, sensual pleasure and relief (Ekman, personal communication). The main focus of this thesis is on non-verbal vocalisations of emotions, but it examines both verbal and non-verbal expressions of these positive emotions. More recently, Ekman has suggested a more extensive set of potentially basic positive emotions (Ekman, 2003), but to examine these empirically is beyond the scope of this thesis. The methodology used to study recognition of the positive vocalisations was the forced-choice paradigm, which has been commonly used with both facial (e.g., Ekman et al., 1969; Young el al., 1997) and vocal emotional stimuli (e.g., Banse & Scherer, 1996; Laukka, 2004; 2005; Schröder, 2003; von Bezooijen, Otto & Heenan, 1983). This paradigm is simple for participants to understand and perform, and has been found to yield robust results (Ekman, 1994).

Are there vocal equivalents of the facial expressions of the established basic emotions?

The hypothesis has been put forward (Ekman, 1992b; Tomkins, 1962) that there are vocal correlates of the basic emotions anger, disgust, fear, happiness, sadness, and surprise. A few studies have gone some way towards testing this. Scherer, Banse, and Wallbott (2001) tested recognition of pseudo-speech expressing anger, fear, sadness, and joy in university students from the USA, Indonesia, and seven European countries. They found that participants from all of the cultures were able to recognize vocalisations of all of the emotions, at a level that exceeded chance. However, this study only included four of the six basic emotions, and used pseudo-speech created from phonemes exclusively taken from European languages, thus confounding differences in culture with differences in language. Non-verbal vocalisations have been suggested to be perceived as more reliable and spontaneous signals of emotion (Johnstone & Scherer, 2000). A study by Scott et al. (1997) investigated the ability of a patient with bilateral damage to identify emotions from verbal and non-verbal vocalisations of the basic emotions. The data from the control participants showed that both verbal and non-verbal vocalisations of emotions were reliably identified by normal listeners. However, surprise was excluded in the verbal condition and thus the whole set of basic emotions was not examined fully. In addition, the data reported for the control participants was limited and did not include data on the types of errors that the participants made. A more extensive investigation of this issue is desirable. This thesis examines the ability of listeners to identify vocal correlates of the basic emotions from both non-verbal and verbal vocalisations in the context of the extended set of positive emotions proposed by Ekman (1992b).

Are non-verbal vocal expressions perceived as discrete or dimensional entities?

As discussed in sections 1.3, there is a lively debate about whether emotions are discrete or dimensional. This thesis aims to explore this issue in the context of non-verbal vocalisations of emotions. In addition to forced-choice decisions, listeners will also make emotional ratings of the stimuli. These rating data are examined to see whether the "accurate" stimulus type is rated higher than other stimulus types in each rating scale, for example, whether anger stimuli are rated higher on the anger scale than are the other stimuli types. Such a pattern would support the discrete accounts of emotions, which holds that stimuli expressing one emotion are perceived as expressing only that emotion. In contrast, the dimensional account would predict that ratings would rely mainly on similarities in terms of arousal and valence (Russell, 1980). The patterns of confusions and errors listeners make are also examined to see whether these map onto the suggested underlying dimensions. Principal Components Analysis (PCA) is applied to the rating data in order to investigate whether the dimensions of arousal and valence underlie participants' perceptions of these sounds.

Does the stimulus selection procedure affect recognition?

Most previous studies using visual or auditory emotional stimuli have used the best recognized stimuli based on pilot testing and/or experimenter judgment (e.g., Banse & Scherer, 1996; Ekman et al., 1969; Scherer, Banse & Wallbott, 2001; Schröder, 2003). The initial experiments in this thesis used stimulus sets matched for average recognition (see also Scott et al., 1997). In order to examine whether stimulus selection procedures affect listeners' performance, this thesis details comparison of recognition accuracy using a stimuli set selected to match for accuracy versus one selected for best recognition.

Is agreement inflated by the use of the forced-choice paradigm?

As mentioned above, the forced-choice paradigm is commonly used in studies of emotional communication. However, it has been criticized for creating artificially high agreement, as participants are required to select one of the response alternatives offered (Russell, 1994). It has suggested that offering the response alternative "none of the above" would avoid this possible confound (Frank & Stennett, 2001). This methodological manipulation ensures that participants are not forced to choose any of the emotional labels. To ensure that agreement is not artificially inflated by the use of the forced-choice methodology, this thesis includes an examination of the effect of adding the response option "none of the above" to a forced-choice task.

What is the role of different acoustic cues in the recognition of non-verbal expressions of emotions?

Little is known about what acoustic features of non-verbal emotional vocalisations are important for the listener to decode the intended emotional message. One method used successfully with emotional speech stimuli to examine acoustic cues is to degrade the acoustic signal and examine the impact upon emotion recognition (Ladd, Silverman, Tolkmitt, Bergmann & Scherer, 1985; Lieberman & Michaels, 1962). The aim of the investigation in this thesis was to determine experimentally how acoustic factors affect the perceived emotion of non-verbal emotional expressions. In order to examine this, the acoustic structure of the emotional vocalisations was manipulated in a number of different ways, and the effects of these different alterations on participants' ability to identify the emotional sounds were evaluated.

What is the role of different acoustic cues in the recognition of emotional speech and how do they relate to the cues used to recognize emotions from non-verbal expressions of emotions?

Just as with non-verbal vocalisations, emotion recognition from speech may rely on a particular set of acoustic cues. To what extent are these similar or different to those used for non-verbal vocalisations? Several studies have investigated the acoustic basis of the communication of emotions in speech or speech-like utterances (e.g., Banse & Scherer, 1996, Bänziger & Scherer, 2005; Laukka, 2004; Murray & Arnott, 1993). However, these studies have had limited success in establishing specific patterns of acoustic cues that communicate emotional states (Scherer, 1986; Juslin & Scherer, 2005). Many of the acoustic cues that are thought important are difficult to measure (Scherer, 1986), and most studies have examined only a small set of features, mainly relating to pitch (Juslin & Laukka, 2003). Rather than measuring acoustic features, some studies have used acoustically manipulated speech to study the role of different acoustic features in emotional communication (Ladd et al., 1985; Lieberman & Michaels, 1962). The findings from these studies have suggested that several types of cues likely play a role in communicating both emotional states and affective arousal, and that there is unlikely to be a simple relationship between any single cue and specific emotions.

The aim of the investigation in this thesis was to determine experimentally how acoustic factors affect the perceived emotion of emotional speech and to compare these findings with the data collected using non-verbal stimuli. In order to examine this, the acoustic structure of the emotional speech stimuli was altered using the same manipulations as with the non-verbal sounds, to allow for a direct comparison of the effects of these acoustic manipulations on listeners' ability to identify emotions from verbal and non-verbal vocalisations of emotions.

Can acoustic analysis provide sufficient detail to statistically discriminate sounds from different emotional categories?

If the sounds that comprise a given emotional category are acoustically different from other sounds, a statistical model should be able to reliably allot a stimulus to the accurate category. Do sounds produced by human speakers to communicate emotions fulfil this criterion? It is the case that for both facial expressions of emotions and emotional speech, emotion classification can be statistically modelled using basic perceptual cues. For example, research with visual stimuli has shown that statistical methods can successfully discriminate facial expressions of different emotions on the basis of a principal component analysis of pixel intensities (Calder, Burton, Miller, Young, & Akamatsu, 2001). Banse and Scherer (1996) demonstrated that discriminant analysis is also possible for vocal stimuli, using an acoustic analysis of pseudo-speech stimuli expressing 14 different emotions. This thesis extends this work to non-verbal vocalisations of emotions. Measurements of pitch cues, spectral cues and envelope information were used in a discriminant analysis in order to examine whether the information provided sufficient detail to discriminate between the different emotion categories.

Can the acoustic features of emotional sounds predict participants' perception of the sounds, as measured by emotional ratings?

In addition to seeing whether acoustic cues can predict category membership, one aim of this thesis was to examine whether listener's judgments could be predicted from the acoustic features of the sounds. Establishing a role for acoustic features in both production and perception would further our understanding of the communicative nature of non-verbal vocal emotional expressions. What makes us perceive a sound as angrier than another, or more fearful? One of the aims of this thesis is to map out what constellation of acoustic cues is associated with non-verbal vocalisations expressing different positive and negative emotions. Acoustic measurements of the sounds are used in a series of multiple regressions to examine whether they can be used to predict participants' ratings on a set of emotional scales. This allows for an examination of whether a constellation of acoustic cues does predict participants' perceptions of the sounds, as reflected by their ratings on the emotional scales. This is done separately for each emotional scale, and for ratings of arousal and valence of the sounds, to compare the acoustic cues that are associated with each emotion and dimension. Similar studies done with emotional speech or pseudo-speech have found that perceptual measures of emotion, and to some extent arousal, can be predicted from constellations of acoustic cues (Banse & Scherer, 1996; Laukka, 2004).

Can emotions vocalisations be communicated across cultures?

There has not been extensive investigation into the cross-cultural recognition of emotional vocalisations (Juslin & Scherer, 2005). The relatively small number of studies carried out to date has tended to show that emotional vocalisations are identified at a level that exceeds chance in all cultures, although they are recognized less well than emotional facial expressions (Elfenbein and Ambady, 2002b). An important limitation of previous studies is their lack of inclusion of non-Western cultures and pre-literate cultures. To demonstrate cross-cultural consistency more convincingly, studies including non-Western, pre-literate cultures are needed. Norenzayan and Heine (2005) suggest the use of the two-culture approach. This method compares participants from two populations that are maximally different in terms of language, literacy, philosophical traditions etc., and the claim of universality is strengthened to the extent that the same phenomenon is found in both groups. The current thesis applied the two-culture approach to investigate the extent to which emotional vocalisations are cross-culturally recognizable. Importantly, this was done bi-directionally, such that recognition of emotional vocalisations produced in Culture A was studied in Culture B, and vice versa. Work into the cross-cultural recognition of emotional signals has tended to use stimuli produced in one culture (or language group), and study the recognition of those stimuli in a number of different groups, thus neglecting the bi-directionality of emotional communication.

This thesis examined the cross-cultural recognition of non-verbal vocalisations of emotions. All previous studies of emotional signals in the voice have studied emotions in speech (e.g., Beier & Zautra, 1972; van Bezooijen et al., 1983) or pseudo-speech (Scherer, Banse, & Wallbott, 2001). The use of non-verbal vocalisations evades the problem of comparing recognition across groups that vary not only in culture but also in familiarity with the language or phonemes used in the stimuli.

What is the role of spectral detail in speech intelligibility, emotion recognition, and speaker differentiation, and what are the relationships between these processes?

In addition to emotions, the human voice also communicates many other types of para-linguistic information, including cues about the speaker's gender, age, and social class (Karpf, 2006). Despite the wealth of theories on speech-processing, few attempts have been made to create a framework of how different kinds of para-linguistic information in the voice are extracted and processed, and how these processes interact. This was approached using acoustic degradation. As opposed to its use in the instances outlined earlier (see the sections above on "What is the role of different acoustic cues in the recognition of non-verbal expressions of emotions?" and "What is the role of different acoustic cues in the recognition of emotional speech and how do they relate to the cues used to recognize emotions from non-verbal expressions of emotions?"), this manipulation requires a much finer grain of degradation in order to examine the role of spectral information; By gradually varying the degradation of the signal, the importance of this acoustic cue for the different processes can be established. Listeners carry out a number of tasks with the degraded stimuli, identifying the verbal contents, as well as the emotion, and the speaker of the speech segment. This allows a comparison of the extent to which these judgments rely on the same acoustic information.

What are the neural correlates of the perception of non-verbal vocalisations of emotions?

Despite the wealth of data on neural processing of facial signals and emotional speech, little work has investigated the neural underpinnings of non-verbal vocalisations of emotion. One aim of this thesis is to examine how vocal expressions of emotion are processed in the brain, and this is investigated using an fMRI paradigm in Experiment 13. Some claims have been made that there is a "voice selective area" in the Superior Temporal Sulcus (STS; Belin, Zatorre and Ahad, 2002; Belin, Zatorre, Lafaille, Ahad & Pike, 2000), and several studies have also shown this region to be involved in the processing of emotional speech (e.g., Grandjean et al., 2005; Wildgruber et al., 2005). A number of studies of emotional speech have yielded right-lateralised activation (Buchanan et al., 2000; Wildgruber at al., 2005) whereas others have shown bilateral activation (Kotz et al., 2003), and the patterns of activation are therefore hypothesised to be right-lateralised or bilateral, with significant areas of activation in the STS.

Does the perception of non-verbal expressions of emotion activate areas involved in motor planning?

One aim of this thesis was to investigate links between auditory perception and action preparation in the context of emotional vocalisations. Previous studies of emotional facial expressions have shown that passive perception of emotional expressions engage areas involved in motor planning (Carr, Iacoboni, Dubeau, Mazziotta & Lenzi, 2003; Leslie, Johnson-Frey & Grafton, 2004) and studies of speech have shown that auditory stimuli can also elicit activation in areas associated with motor functions (Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Wilson, Saygin, Sereno, & Iacoboni, 2004). These findings suggest a close neural link between the perception of important social signals and the preparation for responsive actions. Experiment 13 aimed to examine whether the link between perception and action previously found for speech and and facial expressions of emotions would also exist for non-verbal vocal expressions of emotions. Based on previous work, it was hypothesized that areas involved in motor planning, such as the left anterior insula (Dronkers, 1996), the inferior frontal gyrus (Carr et al., 2003), and the pre-SMA (Krolak-Salmon et al., 2006) would be engaged in the processing of non-verbal vocalisations of emotions.

2. CAN WE COMMUNICATE EMOTIONS USING NON-VERBAL VOCALISATIONS OF EMOTIONS?

This chapter consists of a series of experiments investigating recognition of emotions from non-verbal vocal expressions, through the use of forced choice and rating tasks. Experiment 1 tests the recognition of expressions of a set of positive emotions, using non-verbal vocalisations of achievement/triumph, amusement, contentment, pleasure, and relief, across two language groups. The results of this experiment indicate that vocalisations of these positive emotions are correctly identified by naïve participants. Experiment 2 extends the set of emotions to include the "basic" emotions anger, disgust, fear, sadness and surprise. Listeners accurately identify all ten classes of non-verbal vocalisations. The rating data are used in a principal components analysis which identifies two underlying dimensions correlating with participants' valence and arousal ratings. Experiment 3 is a replication of Experiment 2, using stimuli selected for best recognition, and excluding the category contentment, as these stimuli were not well recognised in Experiments 1 and 2. Participants are significantly better at classifying stimuli in Experiment 3 as compared to Experiment 2, showing that stimulus selection procedure is important. In Experiment 4 the response option "none of the above" is added to the ten emotion labels used in Experiment 2. Participants remain able to reliably identify emotional vocalisations also when not forced to respond with any of the emotion labels offered, both in overall performance and for each emotion individually. These experiments demonstrate that non-verbal vocalisations are reliable communicative tokens from which naïve participants can infer emotional states.

More than one kind of happiness

As described in Chapter 1, research into vocal expressions of emotions has established that there are recognisable vocal correlates of the "basic" emotions anger, fear, disgust, happiness, sadness, and surprise (e.g., Juslin & Laukka, 2003; Murray & Arnott, 1993; Scherer, Banse & Wallbott, 2001; Scherer, 2003; Scott et al., 1997). However, a number of studies have found that recognition rates for "happy" vocalisations tend to be relatively low (e.g., Banse & Scherer, 1996; Juslin & Laukka, 2001; Scherer, Banse & Wallbott, 2001, study 2). In a cross-cultural study on vocalisations of the basic emotions using speech-like utterances, recognition rates for participants from all of the nine cultures was lowest for happiness/joy expressions (Scherer, Banse & Wallbott 2001). In a recent meta-analysis of emotion decoding including 60 experiments of vocal expressions of emotion and 13 of music performance (Juslin & Laukka, 2003), happiness was one of the least well recognised emotions from both vocalisations and musical segments.

This contrasts with findings from research on emotional expressions in the face, where it is typical to find higher agreement in judgments of happy facial expressions than other emotions, and this applies across cultures (Ekman, 1994; Elfenbein & Ambady, 2002b). This pattern of findings has led Ekman to suggest that happiness could usefully be fractionated into a set of distinct positive emotions (1992b; 2003). Each of these would have a distinct, recognisable vocal signal, but would share the facial expressions of a smile. In the same way that we divide "unhappiness" into anger, fear, disgust etc, we could also divide happiness into discrete positive emotion categories. Indeed, "happiness" may be no more useful a category than "unhappiness", and low accuracy rates may thus be explained by the production of sounds "at the wrong level of description". Instead, vocal expressions of more specific states may be easier to identify; Expressions of 'unhappiness' would likely be difficult to recognise, as they would be expressing too broad a state. Of course, expressions of emotional states that are too narrow may also be difficult to identify, for example, distinguishing between jealousy and envy (Sabini & Silver, 1997).

The dimensional approach

In contrast to Ekman's view that emotions constitute discrete categories, other theorists have proposed that emotions are more fluid entities that blend into each other. Their relations are determined by certain parameters, usually according to where they fall on a number of dimensions. Russell's influential circumplex model (Russell, 1980) characterises emotions in terms of their arousal and valence. In a series of studies specifically investigating whether these dimensions are perceived in vocal expressions, Bachorowski and colleagues found support for Russell's dimensional account (Bachorowski & Braaten, 1994; Bachorowski & Owren, 1995). Bachorowski suggests that rather than the listener inferring specific emotions from others' vocal expressions, "speech acoustics provide an external cue to the level of non-specific arousal associated with emotional processes" (Bachorowski, 1999, p. 55). According to this view, no discrete emotions would be communicated in the voice, but the listener would interpret the sender's emotional state from contextual cues and the arousal state that their vocal signals communicate.

However, several studies using morphed images have provided evidence that challenges the dimensional view by demonstrating categorical perception of facial expressions of emotions (Calder et al., 1996; Etcoff & Magee, 1992; Young et al., 1997). This has been interpreted as supporting a categorical view of emotions, as it indicates that participants perceive emotional expressions as signaling one distinct emotion category at any given time. Young et al. (1997) point out that according to dimensional models, participants would be expected to perceive some morphs as neutral or belonging to a different emotion (see Figure 1.1 of Russell's circumplex model in Chapter 1). For example, according to the circumplex model, a morph between a happy and sad expression would be expected to be perceived as neutral, whereas a morph between a happy and an angry expression would pass through a fearful expression. However, Young et al. found that participants tended to identify stimuli as belonging to one of the end-point categories, with few intrusions from other categories, including neutral. The authors concluded that "these results provide strong evidence that facial expressions are perceived as belonging to discrete categories" (p. 308).

The categorical perception paradigm has recently been applied to auditory stimuli: Laukka (2005) created blends of emotions using synthesized stimuli. He measured the fundamental frequency, temporal cues, voice intensity, and spectral energy distribution from speech expressing different emotions. By incorporating acoustic elements from the emotional sounds into a neutral sentence, he was able to produce synthesized emotional expressions. These were based on a neutral expression, so that properties that were not manipulated, such as formant structure, would remain neutral. Laukka examined the perception of blends of two emotions, using a match-to-sample task and a forced choice task. He found evidence for categorical perception in vocal expressions of emotions from both tasks. In a more extensive discussion of the issue of dimensional aspects of emotions, Laukka (2004) suggests that that although data indicates that emotion dimensions may not correspond to perceptual aspects of emotions, dimensions may constitute cognitive elements of affective function.

The approach used in this chapter

This chapter investigates emotion recognition in non-verbal vocal stimuli. Non-verbal vocalisations are sounds that do not convey linguistic information. Most research into emotional vocalisations has studied emotional prosody in actual speech or speech-like sounds. Here, the focus is on non-verbal communicative tokens such as laughter, sighs and screams. The main issue I wish to examine is whether these kinds of sounds reliably communicate emotional meaning: Can naïve participants identify the emotions expressed in these brief, non-verbal vocalisations? In cases where they cannot, what kinds of errors do they make?

One approach that has been previously used to compare categorical and dimensional accounts is categorical perception. However, given the multi-dimensionality of the acoustics of human vocalisations in general and non-verbal vocalisations of emotions in particular, it is difficult to create stimuli that take into account all the acoustic features of the original sounds. This thesis instead uses an alternative way of comparing categorical and dimensional accounts, by examining participants' ratings of emotional stimuli. By asking participants to rate a set of stimuli on a number of emotion dimensions it is possible to test whether their ratings follow a categorical or a dimensional pattern. According to the categorical account, only the stimulus type corresponding to each rating scale would be rated higher and all other stimulus types would be rated lower on that scale. In contrast, the dimensional account would predict that emotions would blend into each other, and so ratings on each scale would vary gradually between stimulus types. Specifically, Russell's circumplex model would also predict that the extent to which stimulus types were rated highly on other emotional scales would correspond to how close they were in arousal and valence to that emotion.

The aims of the current experiments

This chapter sets out to test Ekman's hypothesis that there is a set of positive emotions with distinct, recognisable vocal signals (Ekman 1992b; 2003). Categorisation and rating tasks are used to examine whether these sounds communicate distinct emotional states, as suggested by Ekman (1992b), or whether they are fluid entities better characterized in terms of their arousal and valence, as suggested by Russell (1980).

Specifically, Experiment 1 tests the hypothesis that the positive emotions achievement/triumph, amusement, contentment, sensual pleasure and relief (Ekman, personal communication) can be reliably recognised and rated from non-verbal vocal expressions of emotions. This experiment is repeated in a second language group to test its reliability. The second experiment investigates both positive and negative emotions, extending the set from Experiment 1 to include non-verbal vocal signals of anger, disgust, fear, sadness, and surprise, in identification and rating tasks. The rating data are used to investigate potential contributions of dimensional and categorical factors using principal components analysis. Experiment 3 compared recognition rates using stimuli selected for best recognition based on pilot data, with recognition rates using stimuli matched for recognition rates (Experiment 2). Experiment 3 also excluded the category contentment. In Experiment 4 the response option "none of the above" is added to the ten emotion labels used in Experiment 2, to investigate the possibility of inflated agreement in Experiments 1–3 due to the use of forced choice tasks.

Experiment 1a — Recognition of non-verbal vocalisations of positive emotions

Before performing a full analysis of a range of positive and negative emotional vocalisations, the claim that there are distinct communicative signals for the proposed positive emotions needed to be tested (Ekman, 1992b). In this first experiment this was done by investigating whether naïve listeners could identify candidate vocal expressions correctly and whether their rating of the sounds would be consistent.

Method

Stimulus preparation and pilot

The non-verbal expressions of emotion were collected from two male and two female native British English speakers. None of these speakers were trained actors. Speakers were recorded in an anechoic chamber (a soundproof room with no reverberation) and were presented with appropriate scenarios for each emotion label (see Appendix A). These scenarios were composed by the experimenter and aimed to describe situations that would elicit each of the relevant emotions. No explicit guidance was given as to the precise sort of sounds the speakers should generate, that is, the speakers were not given exemplars to mimic (to avoid artifactual stimulus consistency). Speakers were asked to produce a variety of sounds. Most importantly, they were instructed not to produce 'verbal' items (e.g., 'phew!', 'yippee!'). Each speaker produced at least 15 sounds per category. The resultant 240 sounds were digitised at 32kHz.

All the stimuli were then piloted on 10 participants, who performed a forced-choice task that was procedurally identical to the main study (see below). This method was used to identify and remove the stimuli that were least well recognised; many such stimuli were due to poor production, as the speakers often found it difficult initially to produce some classes of stimuli on command (e.g., sensual pleasure), and also spent some time trying different sounds for other conditions (e.g., achievement/triumph). The pre-selection of stimuli based on the results of pilot tests is commonly performed in emotional expression studies (e.g. Banse & Scherer, 1996; Schröder, 2003) to avoid experimenter bias that would arise from a more subjective stimulus selection procedure (though both Banse & Scherer (1996) and Schröder (2003) used initial expert judgments of the stimuli). A test set was chosen from the results of the pilot on the basis of the recognition scores for each stimulus. To aim for even stimulus recognition standard, 16 tokens were chosen for each category, with an average inter-judge agreement of 78% across all categories. This technique was used to control for differences between emotion conditions that arise from uneven stimulus generation — for example, if the best recognised stimuli for an emotion X were recognised at 99%, while the best recognised stimuli for emotion Y were recognised at 68%, subsequent comparisons of the two stimulus classes (e.g. when the acoustic structure was distorted — see Chapter 3) would likely yield significant differences that were an artifact of the original recognition rates. Differences between the standards of recognition across emotional conditions can arise from difficulties speakers experience when producing the sounds (see above), and matching recognition rates at a level below ceiling allows for a comparison across emotional conditions while trying to control for such variation. All speakers were represented in each set of stimuli for each emotion, with the exception of male speaker 1 for anger, and male speaker 2 for sensual pleasure. Examples of the stimuli are available on a CD in the additional materials.

Participants

Twenty British English speaking participants (10 male, mean age 28.2 years) were tested. The participants were recruited from the University College London Psychology department participant database.

Design & Procedure

Categorisation task The forced-choice categorisation task consisted of assigning a label to each emotional sound. The labels were visible throughout testing. Each label was introduced alongside a brief emotion scenario (see Appendix A) in the instruction phase, and the response options were presented in alphabetical order.

Rating tasks Each rating task consisted of judging the extent to which each stimulus expressed the given dimension on a 7-step scale, with 1 denoting the minimum and 7 the maximum. There were seven rating tasks, one for each positive emotion and in addition scales for arousal (minimal-maximal), and valence (negative-positive).

Testing All participants carried out the categorisation task first, and then completed the rating tasks in a random order. Each stimulus was played through headphones from a lap top computer using the Psyscope program (Cohen, McWhinney, Flatt, & Provost, 1993). The response was given as a key press on the numbered keys, with each of the numbers 1–5 representing each of the emotion labels in the categorisation task, and using the numbers 1–7 in the rating tasks. In the categorisation task, the labels were accompanied by the emotion scenarios to aid understanding. The response options were visible in alphabetical order on the screen, and the labels and scenarios were available on a sheet of paper in front of the participant throughout the testing session. These scenarios gave an example of a situation eliciting that emotion, and were the same sentences as those used to elicit the stimuli (see Appendix A). These sentences were also used as examples in the rating tasks for the scales based on those emotions. Two contrasting scenarios were given each for arousal and valence, e.g., minimally aroused and maximally aroused, as the absence as well as presence of these features is distinctive. Having rated all of the stimuli on one scale, the participant then rated all of the stimuli on the next scale, thus hearing all of the stimuli a total of eight times, once for each of the rating tasks and once in the categorisation task. Within each task, the 80 (16 for each emotion) stimuli were played in a random order.

Results

The listeners categorised the positive emotion sounds accurately (see Table 2.1 A), that is, for each stimulus type the most frequent response was the appropriate category. Due to technical problems there was a small number of missing data points across all participants (18 responses missing). Chi-square analyses of the categorisation data revealed that participants were significantly above chance for all stimulus categories. The chi-values for achievement/triumph, amusement, contentment, sensual pleasure and relief were $\chi_{(4)}$ = 924.0, 935.7, 244.4, 435.1, 818.4, all $p < 0.0001$. Proportions of correct categorisations ranged between 52.4% (contentment) and 90.4% (amusement), where chance would fall at 20%. There was some systematic confusion between the categories contentment

Tab. 2.1: Categorisation of positive emotion vocalisations by English (A) and Swedish (B) participants (%). Correct categorisations are given in bold type; horizontal rows add to 100.

Stimulus type	Response				
	Achievement	Amusement	Content	Pleasure	Relief
A		England (n = 20)			
Achievement	**88.4**	4.7	1.9	1.9	3.2
Amusement	1.9	**90.4**	1.6	3.9	2.3
Contentment	7.9	5.0	**52.4**	25.2	9.5
Pleasure	0.3	0.4	29.9	**61.6**	7.9
Relief	0.3	0.3	10.1	5.3	**83.9**
B		Sweden (n = 20)			
Achievement	**70.9**	14.1	4.5	1.9	8.8
Amusement	2.5	**80.6**	4.1	5.3	7.2
Contentment	8.8	2.8	**47.8**	26.9	12.5
Pleasure	0.9	1.3	32.8	**56.9**	8.1
Relief	5.3	0.3	13.1	13.1	**67.8**

and sensual pleasure, with over 29% of sensual pleasure sounds being categorised as contentment, and participants categorizing 25% of contentment sounds as sensual pleasure.

On each scale, the participants rated the correct emotion highest, with the exception of contentment sounds (see Table 2.2A). These ratings were tested with a repeated measures Analysis of Variance (ANOVA) for each emotional rating scale, and the use of planned comparisons. Every ANOVA was significant ($F_{(4,76)}$ = 75.2 for achievement/triumph, 87.0 for amusement, 8.9 for contentment, 7.2 for sensual pleasure and 13.6 for relief, all $p < 0.0001$), evidence that for each of the rating scales, participants' ratings varied across stimulus types. To test whether the "correct" stimulus type for each scale was significantly more highly rated than the mean of the other emotional classes, planned comparisons were performed for each ANOVA. These were significant for each emotional rating scale ($t_{(19)}$ = 17.7 for achievement/triumph, 11.7 for amusement, 2.3 for contentment, 3.8 for sensual pleasure and 3.4 for relief, all $p < 0.05$). ANOVAs were also performed on the ratings for arousal and valence. The results indicated that there was significant variation across both scales with emotional stimulus condition ($F_{(4,76)}$ = 19.8 for arousal and 61.9 for valence, both $p < 0.0001$). No prior hypothesis existed about these patterns, so no planned comparisons were performed. Descriptively, the valence and arousal ratings were highest for achievement/triumph, also high for amusement sounds and lowest for relief.

Discussion

This experiment shows that participants consistently categorise and rate non-verbal vocal expressions of positive emotions, providing initial support for Ekman's hypothesis of a set of distinct

Tab. 2.2: Ratings of positive emotion vocalisations by English (A) and Swedish (B) participants, (min = 1, max = 7). Ach = Achievement/Triumph, Amu = Amusement, Cont = Contentment.

Stimulus type	Response						
	Ach	Amu	Cont	Pleasure	Relief	Arousal	Valence
A				England (n=20)			
Achievement	**6.4**	4.3	4.3	4.0	5.0	5.5	6.3
Amusement	4.0	**5.9**	4.0	4.2	3.8	4.6	5.6
Contentment	3.2	2.4	**4.7**	4.2	3.5	3.3	4.2
Pleasure	3.1	2.6	5.1	**5.1**	3.5	3.6	4.7
Relief	2.9	2.0	3.5	3.1	**5.2**	2.6	3.5
B				Sweden (n=20)			
Achievement	**6.0**	5.2	4.8	3.5	5.3	5.9	5.9
Amusement	3.4	**5.6**	4.7	4.0	4.3	4.8	5.6
Contentment	2.6	2.5	**4.6**	4.6	3.2	2.7	4.4
Pleasure	2.3	2.7	5.0	**5.7**	3.4	2.5	4.6
Relief	2.4	1.9	3.2	3.6	**4.9**	2.2	3.5

vocal expressions for positive emotions (Ekman, 1992b). Not only were listeners able to identify emotional sounds at a level that reliably exceeded chance, but they also rated each class of positive emotion accurately. It is noteworthy that the data from both the categorisation and rating tasks suggest contentment to be the weakest of the positive emotions. It is possible that this emotion reflects a subset of sensual pleasure, and does not constitute a separate emotion category. In addition, this weakness could be due to contentment being an emotion of relatively low intensity; previous studies have found that vocal emotions of stronger emotion intensity were easier to decode than those of weaker emotion intensity (Juslin & Laukka, 2001). This experiment did not directly study the intensity of the stimuli, but contentment was rated as being relatively low in arousal, which could be a related feature.

Could the consistency in the participants' responses have arisen solely from an original bias in the way the expressions were elicited? This is unlikely, as the vocal expressions themselves were not instructed or copied, and there was substantial variation in the vocal expressions that were generated. In summary, this experiment identified five positive non-verbal expressions of emotion that can be reliably recognised. Amusement, achievement/triumph, sensual pleasure, relief and contentment could be reliably recognised and rated from non-verbal vocal expressions. Participants were highly successful at categorizing the stimuli, with performance significantly above chance for all stimulus types. As predicted, participants rated every stimulus type highest on its own scale. The data from this experiment suggest that the vocal emotion category 'happiness' may be better characterized as several positive emotions, and yields the first empirical support for Ekman's (1992b) hypothesis of several vocally expressed positive 'basic' emotions.

Experiment 1b — Recognition of positive emotional vocalisations in a non-English sample

This experiment investigated the reliability of the results from Experiment 1a by replicating it in a different language group. While some attenuation of performance might be expected in a group of listeners from a different language group than the stimulus producers (Elfenbein & Ambady, 2002a) this experiment aimed to establish whether the non-English listeners would recognise these stimuli at rates above chance. Likewise, it sought to establish whether the confusions and ratings of the sounds would be similar between language groups.

Method

Stimulus preparation and pilot

The same stimulus set as in Experiment 1a was used.

Participants

Twenty Swedish participants (10 male, mean age 39.5 years) were tested with instructions in Swedish. The participants were from Stockholm, Sweden. Due to a technical problem, one participant did not complete one of the rating tasks (for the sensual pleasure scale).

Design & Procedure

The design and procedure were identical to that used in Experiment 1a. The labels in Swedish were: achievement/triumph — prestation, amusement — munterhet, contentment — nöjd, pleasure — njutning, relief — lättnad, valence — negativitet-positivitet, arousal — energi (minimal-maximal).

Results

The Swedish listeners categorised the positive emotion sounds accurately (see Table 2.1B). Due to technical problems there was a small number of missing data points (6 missing responses). Chi-square analyses of the categorisation data revealed that the Swedish participants' performance was significantly above chance for all stimulus types. The chi-values for achievement/triumph, amusement, contentment, sensual pleasure and relief were $\chi_{(4)} = 532.7$, 738.5.8, 207.8, 380.8 and 477.4, all $p < 0.0001$, respectively. Proportions of correct categorisations ranged between 47.8% (contentment) and 80.6% (amusement). As in the English sample, there was some systematic confusion between the categories contentment and sensual pleasure, with over 32% of sensual pleasure sounds being categorised as contentment, and 26% of contentment sounds as sensual pleasure. Descriptively, the pattern of ratings was highly similar to that of the English sample. On each scale, the Swedish participants rated the correct class of positive emotion sound highest, with the exception of contentment and relief sounds (see Table 2.2B). When rated for contentment, sounds of sensual pleasure, achievement/triumph and amusement were rated higher than contentment sounds. The Swedish participants rated achievement/triumph sounds as higher on the relief scale than the relief sounds. As before, the ratings were tested with repeated measures ANOVA for each emotional rating condition and the use of planned comparisons. As for the English participants, every ANOVA was significant ($F_{(4,76)} = 69.8$ for achievement/triumph, 148.5 for amusement, 11.8 for contentment, 14.1 for sensual pleasure and 16.4 for relief, all $p < 0.0001$), and the planned comparisons were significant for each emotional rating scale, except for the ratings on the contentment scale ($t_{(19)} = 13.5$ for achievement/triumph, 12.1 for amusement, 5.6 for sensual pleasure and 2.0 for relief, all $p < 0.01$, and $t_{(18)} = 6.3$ for sensual pleasure, all $p < 0.0001$). Similarly to the English sample, ratings for arousal and valence were highest for achievement/triumph, also high for amusement sounds and lowest for relief, and there was significant variation in the ratings between stimulus types ($F_{(4,76)} = 160.4$ for arousal and 40.0 for valence, both $p < 0.0001$).

Comparing Swedish and British participants' performance

To compare performance on the categorisation task between the English and Swedish participants, an ANOVA was carried out on the correct categorisation scores. Language group was a between-groups factor, and emotion category was a within-subject factor. There was a main effect of emotion ($F_{(1,38)} = 15.5$, $p < 0.001$), and a main effect of language group ($F_{(1,38)} = 8.3$, $p < 0.01$), but no significant interaction between the variables was found.

An ANOVA was carried out of the rating scores from the two language groups. The language group was a between-group factor, and stimulus type and rating scale were within-subject factors. For the rating scores there was no main effect of language group, but main effects of stimulus type ($F_{(4,148)} = 103.3$, $p < 0.0001$) and of scale ($F_{(6,222)} = 26.3$, $p < 0.0001$) were found. There was no interaction between language group and stimulus type, or between language group and rating. There was a significant interaction between scale and stimulus type ($F_{(24,888)} = 52.5$, $p < 0.0001$), indicating that stimuli from the different stimulus types were rated differently on the scales. There was also a significant 3-way interaction between the group, scale and stimulus type ($F_{(24,888)} = 2.5$, $p < 0.0001$). In the absence of a significant main effect of language group, such complex interactions are hard to interpret with respect to linguistic differences. However, this could be reflecting the finding that in the English group, each of the stimulus types were rated higher on its own scale than the mean of the other scales, whereas contentment stimuli were not rated higher than the other stimuli in the Swedish group.

Discussion

The results of this experiment indicate that there is consistent categorisation and ratings of the vocal expressions across two language groups. This strengthens the support for Ekman's hypothesis of a set of distinct vocal expressions of positive emotions (Ekman, 1992b). Not only were listeners from both groups able to identify emotional sounds at a level that reliably exceeded chance, but the pattern of their performance on both the categorisation and the rating task was very similar. It is also noteworthy that the data from both samples suggest contentment to be the weakest of the positive emotions.

Clearly, the basic emotion argument would be weakened by large inter-language group differences: Crucially, the two language groups tested showed very similar response patterns. However, they did differ in their ability to correctly categorise the stimuli, with the accuracy of the Swedish group being lower. This kind of cultural advantage has been reported in several meta-analyses (Elfenbein & Ambady, 2002b; Juslin & Laukka, 2003), and has been proposed to be the result of subtle cross-cultural differences (Elfenbein & Ambady, 2003). However, since this study did not include production as well as decoding in both cultures the implications of this study for the issue of cross-cultural differences on emotion communication are limited (Matsumoto, 2002). In addition, this data set does not support a basic emotion account over other accounts of emotional structure; it is simply consistent with Ekman's predictions derived from a basic emotion perspective (Ekman, 1992b).

In summary, amusement, achievement/triumph, sensual pleasure, relief and contentment could be reliably recognised and rated from non-verbal vocal expressions in two language groups. This is strong initial data to support the claim that "happiness" is an emotional category which can be usefully fractionated into several positive emotions with distinct, recognisable vocal expressions (Ekman 1992b).

Experiment 2 — Categorisations and ratings of positive and negative expressions of emotion

Previous work has used non-verbal emotional vocalisations of the 'basic' six facial expressions of emotion (Scott et al., 1997). Such non-verbal vocalisations are arguably better analogues than verbal stimuli to the facial expressions commonly used, as non-verbal noises, like faces, contain no meaningful lexical information. In this experiment, the perception of these vocal expressions of emotions is investigated together with the putative positive emotions outlined in Experiment 1. In addition, including a range of positive and negative emotions allows for an investigation into the contributions of categorical and dimensional factors in the ratings of these emotional signals, by the application of principal components analysis.

Method

Participants

Twenty participants (10 males, mean age 21.4) took part in the categorisation experiment and another 20 (11 males, mean age 25. 5) in the rating tasks. All participants were recruited from the University College London Psychology participant database. None had participated in Experiment 1 or any of the pilot studies.

Stimuli

The same four speakers used in Experiment 1 were recorded using an identical procedure, producing non-verbal vocal stimuli for the emotional conditions fear, surprise, anger, disgust and sadness. As previously, the speakers were not instructed in how these should be produced, and no exemplars were given (see Appendix A for scenarios used). As before, the stimuli were piloted on 10 participants, to remove the poorest exemplars from the stimulus set.

The stimuli used in the categorisation and rating tasks were 100 non-verbal emotion sounds with equal numbers expressing each of the emotions achievement/triumph, amusement, anger, contentment, disgust, fear, sensual pleasure, relief, sadness and surprise. The positive stimuli were a sub-set of the stimuli in Experiment 1. The pilot data were used to ensure that the identification rate mirrored that of the stimuli from Experiment 1. Examples of the stimuli are available on a CD in the additional materials.

Design & Procedure

Categorisation task The forced-choice categorisation task consisted of assigning a label to each emotional sound, choosing between the 10 available options, corresponding to the ten emotions. Each label was accompanied by a scenario in the instruction phase (see Appendix A) and the options were presented in alphabetical order.

Rating tasks There were 12 rating tasks, one for each of the ten emotions and two additional scales for arousal (minimal-maximal) and valence (negative-positive). Each rating task consisted of judging the extent to which each stimulus expressed the given dimension on a 7-step scale, with 1 denoting the minimum and 7 the maximum.

Testing As in Experiment 1, each stimulus was played through headphones from a lap top computer using a Psyscope program (Cohen et al., 1993). The response was given as a key press on the numbered keys, with each of the numbers 0–9 representing each of the emotion labels in the categorisation task, and using keys 1-7 in the rating tasks. In the categorisation task, each label was accompanied by a sentence to aid understanding. This set of sentences was an extension of that used in Experiment 1 (see Appendix A). The same sentences were used as examples in the rating tasks for the scales based on those emotions, in addition to the two contrasting scenarios given each for arousal and valence. Within each task, the stimuli were played in a random order. The participants in the rating tasks were given the scales in a random order and hence heard all the sounds twelve times, once for each rating task. The participants in the categorisation task heard the sounds once in a random order.

Tab. 2.3: Categorisation of positive and negative emotional vocalisations (%). N=20. Horizontal rows add to 100. Ach = Achievement/Triumph, Amu = Amusement, Ang = Anger, Cont = Contentment, Dis = Disgust, Ple = Pleasure, Rel = Relief, Sad = Sadness, Surp = Surprise.

Stimulus Type	Response									
	Ach	Amu	Ang	Cont	Dis	Fear	Ple	Rel	Sad	Surp
Ach	**77.0**	5.0	1.5	3.5	0.5	2.5	1.5	0	0	8.5
Amu	0	**79.5**	0.5	6.0	0	0	7.0	0.5	5.5	1.0
Ang	2.5	1.0	**65.5**	1.5	16.5	9.5	1.0	1.0	1.0	0.5
Cont	4.0	4.0	2.5	**46**	2.0	0	29.0	10.5	1.0	1.0
Dis	0	0.5	2.5	0.5	**93.5**	2.0	0	0	1.0	0
Fear	1.5	13.5	0.5	0.5	1.5	**63**	4.0	1.5	12.0	2.0
Ple	0	1.0	0.5	15.0	2.0	0.5	**65**	10.5	3.5	2.0
Rel	1.0	0	2.0	3.5	2.5	0	2.5	**86**	2.0	0.5
Sad	0	0.5	0.5	5.5	2.0	9.5	6.5	6.5	**69**	0
Surp	1.5	1	4.5	0	14.0	10.0	1.0	13	1.0	**54**

Results

Categorisation

Participants were successful at identifying the sounds, that is, the most commonly selected response was the appropriate one for each emotion. Chi-square analyses of the categorisation data revealed that the participants were significantly better than chance (10%) at matching sounds and labels for each stimulus type, when tested against all other emotions, (see Appendix B). The confusion matrix for the categorisation data is shown in Table 2.3.

Ratings

On each scale, the correct stimulus type were rated most highly, although on the contentment scale the sensual pleasure sounds were rated almost as highly as contentment sounds, and on the surprise scale the achievement/triumph sounds were rated almost as highly as the surprise sounds (see Table 2.4). The achievement/triumph sounds were rated as most positive and the disgust sounds as most negative. The achievement/triumph sounds were rated as the most aroused, and relief and sensual pleasure sounds were rated as the least aroused.

The rating data were examined with repeated measures ANOVAs for each emotional rating condition, with stimulus type as a within-subject factor. Planned comparisons were also carried out for each rating condition. Every ANOVA was significant ($F_{(9,171)}$ = 75.9 for achievement/triumph, 76.4 for amusement, 79.8 for anger, 71.0 for contentment, 89.4 for disgust, 64.6 for fear, 65.9 for

Tab. 2.4: Ratings of positive and negative non-verbal emotional vocalisations. N=20. Ach = Achievement/Triumph, Amu = Amusement, Ang = Anger, Cont = Contentment, Dis = Disgust, Ple = Pleasure, Rel = Relief, Sad = Sadness, Surp = Surprise, Val = Valence, Aro = Arousal. "Correct" ratings are given in bold type.

Stimulus type	Rating scale											
	Ach	Amu	Ang	Cont	Dis	Fear	Ple	Rel	Sad	Surp	Val	Aro
Ach	**6.3**	4.7	1.4	4.5	1.2	1.2	4.2	4.5	1.3	4.3	6.2	6.0
Amu	3.8	**5.6**	1.4	4.0	1.4	1.5	3.8	3.2	1.9	3.2	5.2	4.8
Anger	1.8	1.5	**5.5**	1.6	3.9	2.8	1.6	1.8	2.1	2.0	2.1	5.2
Cont	3.2	2.5	1.5	**5.3**	1.5	1.3	4.7	3.7	1.8	2.1	4.6	2.9
Dis	1.3	1.4	3.0	1.5	**5.9**	1.8	1.4	1.4	1.7	1.9	1.9	4.1
Fear	1.9	2.2	2.0	1.9	3.0	**5.1**	2.0	1.8	3.2	2.8	2.6	4.9
Plea	2.8	2.3	1.3	5.2	1.4	1.5	**5.7**	3.8	2.3	2.3	4.8	2.7
Relief	3.0	1.8	1.9	3.5	1.9	1.7	3.4	**5.3**	2.4	2.2	3.8	2.6
Sad	1.5	1.3	1.5	1.8	2.2	2.7	1.9	1.8	**5.1**	1.8	2.1	3.1
Surp	2.7	2.2	2.8	2.4	3.2	3.2	2.5	3.0	2.1	**4.6**	3.3	4.6

sensual pleasure, 47.6 for relief, 68.8 for sadness and 43.1 for surprise, all $p < 0.0001$), showing that there was significant variation on the emotional rating scales of the different stimulus types. To test whether the "correct" stimulus type for each scale was significantly more highly rated than the mean of the other emotional classes, planned comparisons were performed for each ANOVA. These were significant for each emotional rating scale, ($t_{(19)} = 17.2$ for achievement/triumph, 16.7 for amusement, 14.1 for anger, 11.4 for contentment, 19.4 for disgust, 15.5 for fear, 14.2 for sensual pleasure, 8.8 for relief, 19.5 for sadness and 9.1 for surprise, all $p < 0.0001$). ANOVAs performed on the ratings for arousal and valence indicated that there was significant variation across both scales with emotional stimulus condition ($F_{(9,171)} = 53.3$, for arousal and 77.1 for valence, both $p < 0.0001$).

Arousal and valence

The participants' ratings on the arousal and valence scales are plotted in Figure 2.1. To examine the extent to which listeners' categorisation errors and rating patterns corresponded to the proposed underlying dimensions arousal and valence, this figure was examined to identify emotions that were similar in terms of their ratings on these two dimensions. Two candidate pairs were identified: anger and fear, and sensual pleasure and contentment. Anger and fear were both rated between 2 and 2.5 for valence and around 5 for arousal (see Table 2.4). Contentment and sensual pleasure were both rated around 4.7 for valence and 2.8 for arousal (see Table 2.4). According to Russell's (1980) model, emotions that are similar in terms of arousal and valence would be more likely to be confused with each other, and should also be rated as more similar. This pattern was confirmed

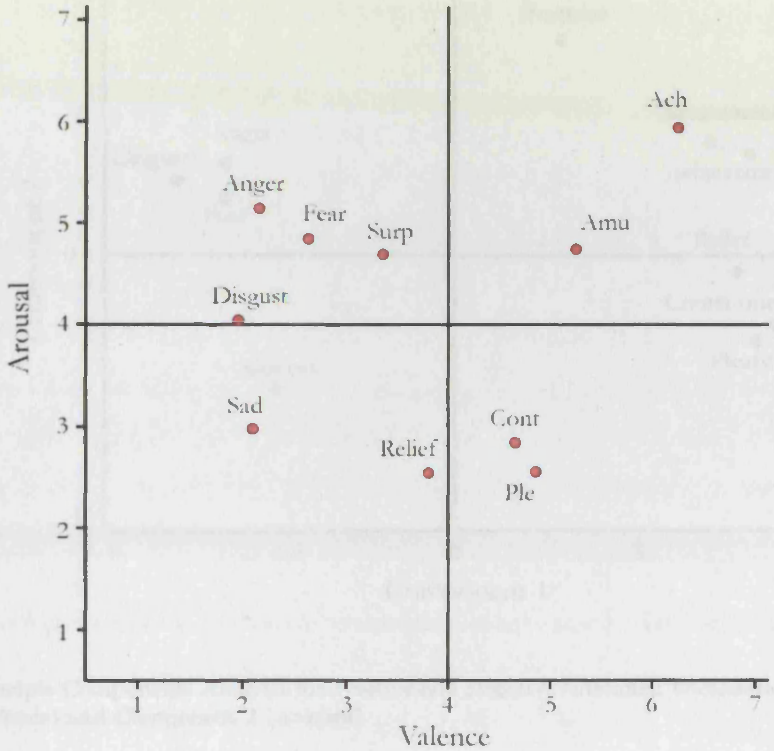

Fig. 2.1: Graphic depiction of participants' ratings of the non-verbal emotional stimuli in terms of valence (horizontal: negative to positive, left to right) and arousal (vertical: low to high, bottom to top).

for contentment and sensual pleasure, where 29% of contentment stimuli were labeled sensual pleasure, and 15% of sensual pleasure stimuli were categorised as contentment (see Table 2.3). In the ratings, contentment stimuli were rated highly on the sensual pleasure scale, and on the contentment scale sensual pleasure stimuli were rated almost as highly as contentment sounds (see Table 2.4). However, anger and fear sounds, although rated similarly on arousal and valence, were not consistently confused or rated highly on each other's scales. Only 0.5% of fear sounds were perceived as communicating anger, and although 9.5% of anger sounds were categorised as fear, they were not the most common confusion for anger sounds. When erroneously labeled, anger sounds were perceived as expressing disgust, an emotion which was rated lower in terms of both arousal and valence (see Table 2.3 and Table 2.4). In addition, anger sounds were not rated highly on the fear scale, and fear sounds were not rated highly on the anger scale. In sum, although the ratings and confusions for contentment and sensual pleasure does seem to fit Russell's model, the data for fear and anger suggests that this is not a consistent pattern. The confusions and ratings do not seem to simply map onto the stimuli's ratings for arousal and valence.

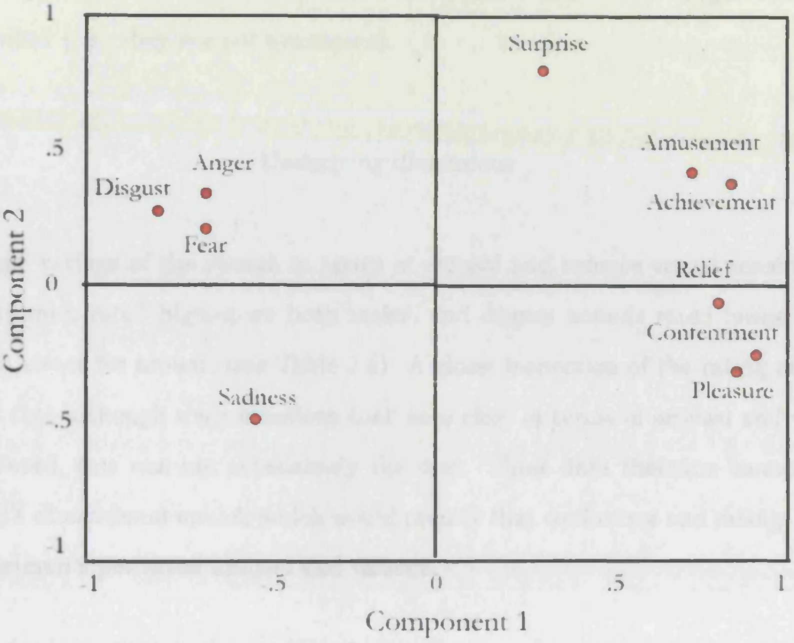

Fig. 2.2: Principle Components Analysis for Positive and Negative emotional Vocalisations. Component 1 (Valence) and Component 2 (Arousal).

Principal Component Analysis

The participants' ratings on all of the scales except arousal and valence were subjected to a principal components analysis (PCA). This procedure examines whether a smaller number of dimensions can account for a significant proportion of the variance in the data. The analysis yielded two factors with eigenvalues over 1, that accounted for 53.3% and 15.7% of the variance of the participants' ratings, respectively. Component 1 correlated with the participants' valence ratings at .97 and Component 2 correlated with the participants' arousal ratings at .87. See Figure 2.2 for a visual representation of the Principal Components Analysis.

Discussion

The results from the current experiment indicate that the emotions fear, anger, disgust, sadness and surprise have clear, recognisable, vocal, non-verbal expressions, and that this can be extended to encompass positive emotions (achievement/triumph, amusement, relief and sensual pleasure). The fifth candidate positive emotion, contentment, seemed to be categorised and rated as a weaker form of sensual pleasure, although contentment and sensual pleasure are arguably not strongly semantically linked (i.e., they are not synonyms).

Underlying dimensions

The participants' ratings of the stimuli in terms of arousal and valence varied considerably, with achievement/triumph rated highest on both scales, and disgust sounds rated lowest for valence, and relief rated lowest for arousal (see Table 2.4). A closer inspection of the rating and confusion data suggested that although some emotions that were close in terms of arousal and valence were commonly confused, this was not consistently the case. These data therefore cannot be said to support Russell's dimensional model, which would predict that confusions and ratings consistently map onto the stimuli's perceived arousal and valence.

The results from the PCA indicated that the two factors, strongly correlating with valence and arousal, accounted for a total of 69% of the variance in the ratings data (see Figure 2.2). Although not identical with listeners' ratings for arousal and valence (see Figure 2.1), the plots from the PCA and the arousal and valence ratings are broadly similar. This confirms that the components of the PCA correspond to the perceived dimensions of arousal and valence, a pattern which would seem to support Russell's model. However, although the PCA plot looks somewhat like a circumplex, this pattern is not reflected in the listeners' actual arousal and valence ratings (see Figure 2.1). These data thus provide only limited support for Russell's model.

The two factors in the PCA were highly unequal in their contribution (53% for "valence" and 16% for "arousal"), with the pattern suggesting a dominant role for valence. This is inconsistent with Bachorowski's (1999) interpretation of dimensional factors in emotional vocalisations, which emphasises the dominant role of arousal. Her work in this area mainly consists of studies analysing acoustic cues in vocal production data. It is possible that the manipulations used in her studies were more successful in eliciting changes in arousal than valence. It may also be that the measurements used in Bachorowski's research relate mainly to the acoustic features involved in signaling arousal rather than valence, or that valence is not a dimensional cue, or at least not one that can be identified acoustically (this point is discussed further in Chapter 3).

In sum, this experiment has identified a set of positive and negative emotional non-verbal vocalisations that can be reliably identified by nave listeners. Participants rated the stimuli consistently, with each stimulus type being rated highest on its own scale. The PCA suggests that arousal and valence may underlie listeners' internal emotional space in the sense of how emotions relate to one another. However, these dimensions do not consistently map onto listeners' categorisation errors and ratings of the stimuli.

Experiment 3 — Recognition of vocal expressions of emotions using the best available stimuli rather than matched stimulus sets.

Experiments 1 and 2 used forced-choice tasks to establish that non-verbal vocalisations of negative and positive emotions can be reliably recognised. In those experiments, stimuli were selected to match for recognition across emotion categories. This stimulus selection strategy controlled for the relative difficulty of stimulus production for the different emotions, and has been used in previous work (Scott et al., 1997). Due to this stimulus quality matching procedure, recognition was not maximized for most categories in Experiments 1 and 2. Many previous studies using both visual and auditory emotional stimuli have used the best recognised stimuli based on pilot testing and/or experimenter judgment (e.g., Banse & Scherer, 1996; Ekman et al., 1969; Scherer, Banse & Wallbott, 2001, Schröder, 2003). Experiment 3 is a replication of Experiment 2, using stimuli with the highest recognition rate, based on pilot data.

This experiment also excludes the category contentment: Both categorisation and rating data from Experiments 1 and 2 revealed contentment stimuli to be the least well recognised and that they were consistently confused with sensual pleasure, suggesting that contentment may not be a distinct emotion category. It is hypothesised that recognition of sensual pleasure stimuli may improve in the absence of contentment sounds, as the two were confused in Experiments 1 and 2.

Method

Participants

Twenty participants (9 males, mean age 29.5 years) from the UCL Psychology Participant Database took part in the experiment. None of the subjects had participated in any other task using emotional vocalisations.

Stimuli

The stimuli used were 90 non-verbal emotion sounds with equal numbers expressing each of the emotions achievement/triumph, amusement, anger, disgust, fear, pleasure, relief, sadness, and surprise. The stimuli were part of the same corpus of stimuli that was produced for Experiments 1 and 2, selected on the basis of the pilot experiments for those experiments. The set used in the current experiment were the stimuli with the highest average recognition scores, with an average of 97.7% correct recognition in the pilot experiment compared to the 78% recognition of the stimulus selected for Experiment 2. Note that mean recognition in Experiment 2 was 72.7%.

Design & Procedure

The design and procedure was identical to that of Experiment 2, except that the keys 1–9 were used as response options in the current experiment. The labels and sentences used were identical to those used in Experiment 2 (see Appendix A).

Results

Participants were highly successful at categorising the sounds (see Table 2.5), with correct responses ranging between 73% (pleasure and sadness) and 94% (disgust). Chi-square analyses of the categorisation data revealed that the participants were significantly better than chance at matching sounds and labels for all categories (see Appendix C).

Tab. 2.5: Confusion data the stimulus set in Experiment 3, selected for best recognition (%). Correct recognition rates in bold. Note that chance level is 11.11%. Horizontal rows add to 100. Ach = Achievement/Triumph, Amu = Amusement, Ang = Anger, Cont = Contentment, Dis = Disgust, Ple = Pleasure, Rel = Relief, Sad = Sadness, Surp = Surprise.

Stimulus type	Response								
	Ach	Amu	Ang	Dis	Fear	Ple	Rel	Sad	Sur
Ach	**81.5**	6.5	1.5	0	1.0	1.0	0	0.5	8.0
Amu	1.0	**88**	0	0	0	1.0	0.5	9.5	0
Ang	2.5	0	**87**	5.0	1.0	1.0	1.0	1.5	1.0
Dis	0	0	2.5	**94**	0.5	0.5	2.0	0	0
Fear	0	0	3.5	1.0	**80**	2.5	0.5	3.0	9.5
Ple	0	0	1.0	1.0	0.5	**73**	12.5	10.5	1.5
Rel	0.5	0	1.0	1.5	0.5	4.5	**89**	0.5	2.5
Sad	0	0.5	0.5	2.5	4.5	10.0	5.5	**73**	3.5
Sur	4.0	0	0	1.0	5.0	1.0	3.0	0	**86**

Tab. 2.6: Difference when stimuli matched for quality (Experiment 2) and when best stimuli used (Experiment 3), calculated as Experiment 3 minus Experiment 2. Data are in percentages.

Emotion	Experiment 2	Experiment 3	Difference
Achievement/Triumph	77.0	81.5	4.5
Amusement	79.5	88.0	8.5
Anger	65.5	87.0	21.5
Disgust	93.5	94.0	0
Fear	63.0	80.0	17
Pleasure	65.0	73.0	8
Relief	86.0	89.0	3
Sadness	69.0	73.0	4
Surprise	54.0	86.0	32
Total	73.4	84.2	10.8

In terms of percentages, recognition rates in the 9-way categorisation task of Experiment 3 were higher than that of the previous 10-way forced-choice task (Experiment 2) in all cases except for disgust, where they were equal (see Table 2.6). The overall recognition in Experiment 3 was 83.5%, over 10% higher than the mean recognition rate of 72.7% in Experiment 2.

However, as the categorisation tasks in Experiments 2 and 3 had different numbers of response options, comparisons between accuracy scores should not be made directly; instead, kappa scores were calculated for the individual subjects' performance for each of the stimulus types. Kappa scores indicate performance levels whilst controlling for the different chance levels. These were calculated using the following equation (Cohen, 1960):

$$\frac{P_o - C}{1 - C}$$

P_o = proportion of observed correct performance

C = chance level of the specific task.

A higher kappa value indicates better performance, with 0 being chance and 1 being perfect performance. In Experiment 3, one calculation with a raw score of zero resulting in a small negative kappa value; this score was converted to a kappa value of zero.

The overall kappa scores were compared between Experiments 2 and 3, using an ANOVA with emotion as a within-subjects factor and experiment as a between-subjects factor. The kappa scores are shown in Figure 2.3. There was a significant main effect of stimulus type ($F_{(8,304)} = 11.6$, $p < 0.0001$), indicating that participants were better at recognizing some stimulus types than others.

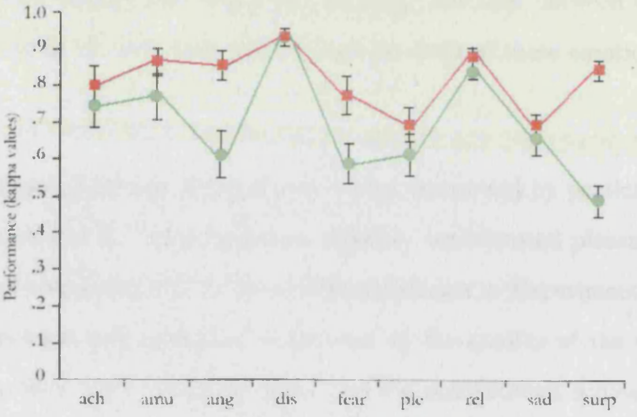

Fig. 2.3: Mean kappa values indicating recognition performance for each emotion condition in Experiments 2 and 3.

There was a main effect of experiment ($F_{(1,38)} = 14.3$, $p < 0.001$), showing that participants were significantly better at categorizing the stimulus set used in Experiment 3 compared to the stimuli used in Experiment 2. There was also a significant interaction between stimulus type and experiment ($F_{(8,304)} = 4.1$, $p < 0.0001$), which reflects the performance pattern illustrated in Figure 2.3.

Using the kappa scores, a series of independent samples t-tests was carried out, in order to compare the performance in Experiment 2 and 3 for each stimulus type. There were significant differences between performance in the two experiments in the anger ($t_{(38)} = 3.2$, $p < 0.001$), fear ($t_{(38)} = 2.5$, $p < 0.05$) and surprise ($t_{(38)} = 7.0$, $p < 0.0001$) conditions. In all of these cases, participants in Experiment 3 performed significantly better than those in Experiment 2.

Discussion

The results from this experiment indicate that selecting the best recognised stimuli from pilot testing leads to higher overall recognition rates. Participants were significantly better at classifying stimuli in Experiment 3 as compared to Experiment 2. However, when examined individually, not all emotions were significantly better recognised in Experiment 3 (although no emotion was less well recognised in Experiment 3 than in Experiment 2). This was reflected by a significant interaction between stimulus type and experiment, (see Figure 2.3). Although participants in Experiment 3 overall performed better than those in Experiment 2, this effect was driven by differences in recognition of certain emotional classes, such as surprise and anger, while recognition rates for other emotions, such as disgust and relief, were virtually identical between the two experiments. The latter is likely due to the fact that performance for both of these emotions were at ceiling in both experiments.

Vocalisations of anger, fear and surprise were better recognised by participants in Experiment 3 than those in Experiment 2. These emotions, together with sensual pleasure and contentment, comprise the less well recognised half of the emotional classes in Experiment 2. From pilot data, contentment was the least well recognised class, and so the quality of the other stimulus types in Experiment 2 was in a sense "dragged down" by the contentment scores. It seems that this matching procedure specifically punished performance for the weaker emotional classes, that is, those types of vocalisations that participants found difficult to identify. For the better recognised emotions such as amusement and disgust, the method used for stimulus selection made little difference. However, the recognition of the weaker emotion classes was significantly improved when the stimuli used were pre-selected for maximum recognisability, indicating that recognition of these classes of stimuli was more sensitive to stimulus quality.

As in Experiments 1 and 2, there was significant variation in Experiment 3 in terms of recognition of stimuli from the different emotion classes, with amusement and disgust sounds being easiest and sensual pleasure and sadness sounds being the most difficult. This confirms that although recognition for all emotional classes was well above chance, some types of vocalisations are more difficult to identify than others.

Sensual pleasure

The stimulus selection procedure did not affect recognition scores for sensual pleasure sounds. It was hypothesised that sensual pleasure sounds would be more easily recognised in Experiment 3, as they had been highly confused with contentment sounds in Experiment 2 and this category was absent in Experiment 3. However, sounds of sensual pleasure were not significantly better recognised in Experiment 3 as compared to Experiment 2, although recognition was somewhat higher (73% in Experiment 3 and 65% in Experiment 2, a difference of 8%). Could it be that this pattern of results reflects participants being reluctant to using the label sensual pleasure? This is unlikely: a closer inspection of the data shows that this is not the case: sensual pleasure comprised 9.5% of participants' responses (use of the labels ranged between 9.0% for achievement/triumph and 11.4% for relief), which is near to the 11.1% that would be expected by chance. Another interpretation of

Tab. 2.7: Recognition rates in studies with non-verbal emotional expressions.

	Study		
	Experiment 3[1]	Schröder, 2003[2]	Scott et al., 1997 (control sample)
Emotion	Stimulus type		
	Non-verbal vocalisations	Affect bursts	Non-verbal vocalisations
Anger	0.9	0.6	0.7
Disgust	0.9	0.9	0.9
Joy/Happiness	0.8	0.9	0.8
Fear	0.8	0.9	0.9
Sadness	0.7	N/A	0.8
Surprise	0.9	N/A	N/A

[1] Joy/Happiness based on average scores of all positive categories (achievement/triumph, amusement, sensual pleasure and relief)

[2] Joy/Happiness based on average scores of all positive categories (admiration, elation and relief). Fear based on startle scores.

this pattern of results is that some of the sensual pleasure vocalisations produced by the speakers were not entirely convincing. Instead of confusing pleasure stimuli with contentment as in Experiment 2, participants made errors by labelling some sensual pleasure stimuli as relief and sadness in Experiment 3. Thus it seems that some of the stimuli were simply not recognisable as sensual pleasure and participants would select other labels regardless of the available response options. Perhaps these stimuli were not specifically perceived to communicate contentment, but rather *not* perceived as expressing sensual pleasure. Nevertheless, the high confusion rates between sensual pleasure and contentment in Experiment 2 suggest that these two emotions are particularly closely linked, at least in terms of the sounds of non-verbal vocalisations used to communicate them.

Recognition of emotional signals in previous work

A comparison with previous studies is not entirely straightforward, as studies have varied greatly in types of stimuli used, emotion categories included, stimulus selection procedures and general methodology. A number of studies have used forced choice tasks with emotional speech stimuli pre-selected for best recognition rates. These will be discussed in more detail in Chapter 5, which focuses specifically on emotional speech.

Two studies have used forced choice tasks with non-verbal vocalisations of emotions (i.e., vocalisations that were not speech or nonsense-speech), Schröder (2003) and Scott et al. (1997). An overview of the recognition for a number of emotions in those studies is shown in Table 2.7.

Schröder investigated the ability of listeners (naïve and expert) to classify and transcribe the emotional content of non-verbal vocalisations of a range of emotions, including anger and disgust.

Most of the emotion categories included in Schröder's study are not commonly used in emotion research, for example boredom and worry. Although there was some variability in the recognition rates across emotions, the overall recognition rate was 81%, comparable to the recognition rate of 83.5% of the current experiment. Scott et al. (1997) used stimuli matched for recognition rates rather than the best available stimuli. Recognition rates for all categories of non-verbal stimuli are within 8% of the current experiment, despite a difference in the number of response alternatives (nine in the current experiment compared with six in Scott et al.).

In terms of specific emotions, recognition rates for most emotions were quite similar across the three studies. The clear exception is anger sounds, which were only correctly classified 61% of the time in Schröder's study, much lower than in the current experiment (87%). This lower recognition rate was also found by Scott et al. (1997), where non-verbal vocalisations of anger were only identified correctly in 72% of presentations, 15% lower than in the current experiment. Recognition rates for anger sounds from both Schröder (2003) and Scott et al. (1997) are closer to Experiment 2 (66%) than Experiment 3 (87%). The large difference in recognition scores for anger stimuli across Experiments 2 and 3 suggests that recognition of anger sounds seems to be particularly sensitive to stimulus quality.

Summary

In sum, this experiment shows that selecting the best recognised stimuli from pilot testing leads to higher overall recognition rates. This effect is stronger for emotional classes that participants have difficulty identifying. Sounds of sensual pleasure, although confused with contentment in Experiment 2, are not significantly better recognised when the response option contentment is removed: sensual pleasure sounds are instead confused with relief and sadness.

Experiment 4 — An evaluation of the forced choice methodology — adding "none of the above" to the list of response options.

The use of a forced-choice methodology is common in emotion research involving facial expressions (e.g., Ekman et al., 1969; Young el al., 1997) and has also been used in work using vocal stimuli (e.g., Laukka, 2005; Schröder, 2003). In a forced choice task, participants are required to categorise stimuli by selecting one of the labels from a list. This paradigm is convenient in that it is simple for participants to understand and perform, straight-forward to score, and tends to yield robust results. However, the forced-choice methodology has been criticized for producing artificially high agreement on what emotion is communicated by a given stimulus, as the available response options are determined by the researcher (Russell, 1994). As the name suggests the forced-choice format forces respondents to select one of the available options, even when the list does not include what would be the participant's preferred label for a given stimulus.

The set of response options offered can have a great impact on participants' performance. A large number of studies have used Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion set (JACFEE; Matsumoto & Ekman, 1988). The intended emotion is the modal response for all of the basic emotion facial expressions in the JACFEE stimulus set, a pattern interpreted to show that emotional facial expressions directly communicate emotional states to the viewer (Biehl et al., 1997). A study by Russell (1993) challenged this interpretation, demonstrating that a majority of participants categorised angry facial expressions from the JACFEE set variously as contempt, disgust, and frustration, when the response options offered did not include the label anger. According to Russell, his study showed that "forced choice can yield anything from random choice to a consensus, even on, from what researchers have generally concluded, the wrong answer" (p.117, Russell, 1994).

In defence of the forced-choice methodology

Russell's interpretation of his findings has been heavily criticised by Ekman (1994), who points out that it is remarkable that participants across so many studies in so many cultures have agreed on the correct label for certain expressions, if, as Russell suggests, the appropriate alternatives were not offered as response options. Ekman suggests that to seriously undermine the finding of universality of facial expressions of emotions, Russell would have to argue that the appropriate label was absent from the response options for all of the emotions studied. According to Ekman, what Russell's (1993) study demonstrated was merely that in the absence of the most appropriate label, participants choose the most similar alternative. This similarity judgement could be based on overlap of muscle movements for the facial expressions (Tomkins & McCarter, 1964) or on semantic similarity, or a combination of features. According to Ekman, this finding is equivalent to removing the response option "yellow" from a colour categorisation task and finding that participants tend to classify the yellow stimuli as orange. This does not show that the stimuli are orange rather than yellow, but merely that participants tend to resort to the nearest available option if the ideal alternative is not offered. Ekman further points out that research using rating scales have yielded similar results to those using forced choice tasks (e.g., Ekman et al., 1987). He agrees with Russell that free labelling is in a sense the best method for studying what emotion participants perceive in a given stimulus, as it imposes no constraints at all. However, this method introduces substantial difficulties in how to categorise the participants' responses, which is especially delicate in cross-cultural comparisons because of problems with translation. Nevertheless, in studies using free labelling, significant agreement has consistently been found, using participants from a number of cultures (Boucher & Carlson, 1980; Izard, 1971; Rosenberg & Ekman, 1994).

Adding the "none-above" option

Some authors have proposed that adding the alternative "none of these terms are correct" to the list of response options is a relatively simple way of avoiding the potentially inflated agreement in forced choice tasks. Frank and Stennett (2001) carried out a study in which they showed that adding the response alternative "none of these terms are correct" to the standard basic emotion labels in a task in which participants were asked to categorise facial expressions of emotions, did not significantly affect performance. However, Frank and Stennett only examined the effect of overall accuracy and did not statistically test the effect of adding the "none" option to the accuracy in identifying each emotion. They did however report the percentages of each response for each of the stimulus types in a table, showing that the effect of including the "none" option ranged between 3–7%. Because the addition of the "none" response alternative may affect performance for different emotion categories differently, it may be informative to examine these effects individually. A version of the "none" option was also used in a study by Haidt & Keltner (1999). American and Indian participants were instructed to provide their own label if they did not find any the emotion terms offered suitable for a given facial expression stimulus. Consistent with Frank & Stennett's study, only a few percent of participants' responses were made up of the "none" answers, and accuracy was comparable to previous studies that had not included this response alternative.

The aim of the current experiment

The current experiment was a methodological manipulation aiming to examine the possibility of an inflated agreement in Experiments 1-3, due to the use of the forced choice paradigm. In this experiment, the response option "none of the above" was added to the ten emotion labels used in

Experiment 2. The participants were instructed that they should choose this option if they did not find any of the emotional labels fit the sound they had just heard. A comparison of participants' performance in the conditions including and excluding this response option, both overall and for each emotion, allows an evaluation of the extent to which the high agreement found in Experiment 2 was due to the use of the forced choice format.

Method

Participants

Twenty participants (10 males, mean age 23.6 years) from the UCL Psychology Participant Database took part in the experiment. None of the subjects had participated in any other task using emotional vocalisations.

Stimuli

The stimuli used were the same set that was used in Experiment 2, 100 sounds with equal numbers expressing each of the emotions achievement/triumph, amusement, anger, contentment, disgust, fear, pleasure, relief, sadness, sensual pleasure, and surprise.

Design & Procedure

The design and procedure was identical to that of Experiment 2, except that the response options included the response alternative "none of the available" in addition to the ten emotional labels. The keys 0–9 were used for the emotion label response options, and the key "n" was used for the "none" option. In all other respects, the labels and sentences used were identical to those used in Experiment 2 (see Appendix A). The instructions were also the same except for the addition of the sentence "If none of these options suit the sound you just heard, press n."

Results

Participants were highly successful at categorising the sounds (see Table 2.8), with correct responses ranging between 45% (sadness) and 91.5% (amusement). Chi-square analyses of the categorisation data revealed that the participants were significantly better than chance at matching sounds and labels for all categories (see Appendix D). The confusion patterns were similar to that of Experiment 2 (see Table 2.3). On average, the response option "none of the above" was selected in 10.5% of the cases. The prevalence of this response varied greatly between emotions, from less than 2% for amusement and sensual pleasure stimuli, up to approximately 20% of the trials for anger, contentment, and sadness stimuli.

Tab. 2.8: Confusion data for the stimulus set in Experiment 4, with stimulus types down and response options along the top - including the response option "none". Data in percentages (%). Ach = Achievement/Triumph, Amu = Amusement, Ang = Anger, Cont = Contentment, Dis = Disgust, Ple = Pleasure, Rel = Relief, Sad = Sadness, Surp = Surprise, Val = Valence, Aro = Arousal. Correct recognition rates in bold. Note that chance level is 9.01%.

Stimulus type	Response										
	Ach	Amu	Ang	Cont	Disg	Fear	Ple	Rel	Sad	Surp	None
Ach	**66.5**	10.5	0.5	1.0	0	0.5	4.0	1.0	0	8.5	7.5
Amu	0	**91.5**	0	0	0.5	0.5	2.5	0	4.5	0	0.5
Ang	0.5	0	**54.5**	0	20.5	1.0	0	1.5	0	0.5	21.5
Cont	3.5	2.5	1.0	**48.5**	2.0	0	17.5	3.5	0.5	0	21.0
Disg	0	0	3.5	0	**90.0**	0	0	0.5	0	0	6.0
Fear	0.5	11.5	0.5	0	0.5	**60.0**	1.0	0.5	11.0	3.5	11.0
Ple	0.5	0	0	27.0	1.5	0	**67.5**	1.0	1.0	0	1.5
Rel	0	0	1.0	8.0	3.0	2.0	3.0	**68.5**	2.5	6.5	5.5
Sad	0	0	0	4.5	1.0	19.5	2.0	6.0	**45.0**	0	22.0
Surp	0.5	0	0.5	0	8.5	13.0	1.0	11.5	0	**56.5**	8.5

The overall recognition was 64.9%, 7.8% lower than the recognition rate of 72.7% in Experiment 2. In terms of specific emotions, recognition rates in the current experiment were lower than in Experiment 2 for achievement/triumph, disgust, fear, relief, and sadness. The opposite pattern was found for amusement, contentment, sensual pleasure, and surprise (see Table 2.9).

Tab. 2.9: Difference between forced-choice tasks including (Experiment 4) and excluding (Experiment 2) the response option "none", calculated as Experiment 2 minus Experiment 4. Data is in percentages.

Emotion	Experiment 2	Experiment 4	Difference
Achievement/Triumph	77.0	66.5	10.5
Amusement	79.5	91.5	-12.0
Anger	65.5	54.5	11.0
Contentment	46.0	48.5	-2.5
Disgust	93.5	90.0	4.0
Fear	63.0	60.0	3.0
Pleasure	65.0	67.5	-2.5
Relief	86.0	68.5	17.5
Sadness	69.0	45.0	24.0
Surprise	54.0	56.5	-2.5
Total	72.7	64.9	7.9

However, as the categorisation tasks in Experiments 4 and 2 had different numbers of response options, kappa scores were calculated for the individual subjects' performance for each of the

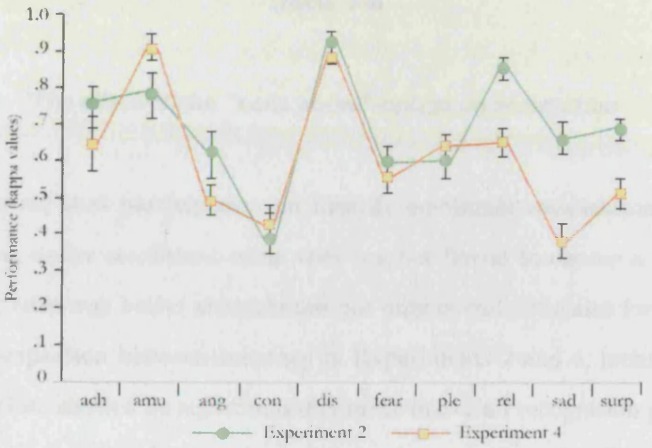

Fig. 2.4: Mean kappa values indicating recognition scores, per emotion condition in Experiments 2 and 4.

emotions. Kappa scores indicate performance levels whilst controlling for the different chance levels (see Experiment 3). A small number of calculations with raw scores of zero resulted in negative kappa values, due to rounding errors inherent in the chance level (1/11=0.090909 recurring). These scores were converted to kappa values of zero.

The overall kappa scores were compared between Experiments 2 and 4 using an ANOVA with stimulus type as a within-participants factor and experiment as a between-participants factor. There was a main effect of stimulus type ($F_{(9,342)} = 23.44$, $p < 0.0001$), indicating that, consistent with Experiments 1–3, participants were better at recognizing some emotional sounds than others. There was no main effect of experiment but there was a significant interaction between stimulus type and experiment ($F_{(9,342)} = 3.4$, $p < 0.001$), which reflects the pattern illustrated in Figure 2.4. This finding indicates that the effect of adding in the "none of the above" option affected the participants' performance to different extents for different stimulus classes.

This was further confirmed in a series of independent samples t-tests, which compared the performance in Experiment 2 and 4 for each emotion condition. The kappa scores for each emotion in Experiments 2 and 4 are shown in Figure 2.4. There were significant differences between the tasks for relief and sadness sounds. In both cases, participants in Experiment 2 performed significantly better than participants in Experiment 4 (for relief $t_{(38)} = 3.6$, $p < 0.001$, for sadness $t_{(38)} = 4.2$, $p < 0.0001$, see Figure 2.4), indicating that the addition of the "none of the above" option reduced the participants' recognition of relief and sadness stimuli.

Discussion

The effect of the "none above" option on recognition

This experiment showed that participants can identify emotional vocalisations at a level that is reliably above chance, under conditions when they are not forced to choose a label. Importantly, participants' performance was better than chance not only overall, but also for each emotion individually. A direct comparison between accuracy in Experiments 2 and 4, including and excluding the "none above" option, showed no significant difference in overall recognition performance. Thus, the availability of the "none above" response option does not significantly affect overall recognition. Although performance was marginally better in the condition where the "none above" label was unavailable, agreement was not significantly inflated by the use of forced choice.

In the current experiment, participants chose the label "none" for 10.5% of the stimuli. This is somewhat higher than in previous work using facial expression stimuli, where participants selected this option in 5–8% of the cases (Frank & Stennett, 2001; Haidt & Keltner, 1999). However, as these studies used visual stimuli and each study included a different set of emotions, comparisons across studies is not straightforward. It may be worth noting that the stimuli used in this experiment were matched for accuracy whereas previous studies employing the "none" option selected stimuli recognised with highest accuracy. Frank & Stennett (2001) included only the basic six emotions using the original Ekman & Friesen (1976) set, whereas Haidt & Keltner (1999) used 14 expressions including embarrassment, compassion and tongue bite, with stimuli selected on the basis of pre-testing.

Specific emotions

The significant interaction between experiment and emotion class indicates that the recognition of certain emotions was affected more than others by the availability of the "none" label. The use of the "none" label varied greatly between emotions, with participants using this option in approximately 20% of the trials for anger, contentment, and sadness stimuli, and for less than 2% for amusement and sensual pleasure stimuli.

The difference in accuracy between the experiments including and excluding the "none" option, were significant only for relief (86 and 68.5% correct, respectively) and sadness sounds (69 and 45%

correct, respectively). In both cases, participants were more accurate at identifying the sounds when the "none" label was not available, indicating that accuracy for these emotions was inflated in the forced-choice condition. However, in both cases the participants' performance was significantly better than chance also in the condition which included the "none" option. The reason for the drop in recognition of relief and sadness sounds may have been inadequacy of the stimuli for these particular emotions (the stimuli used in this experiment were matched for accuracy rather than selected for best accuracy). It is also possible that these stimuli were unique in that they were less similar to sounds of any other category, so that if participants did not perceive them as expressing the correct emotion, they were less likely to categorise them as another emotion.

As already mentioned, comparisons with previous studies are complicated by the variation in the emotions included and the stimulus sets used. Additionally, Frank & Stennett (2001) did not statistically test for an effect of the inclusion of the "none" label for the performance of each of the emotions individually, but only demonstrated the lack of an effect on overall performance. Haidt & Keltner (1999) did not include a condition in which the "none above" option was *not* available and so did not examine the effect of the presence or absence of the label, but rather examined the modal response option for each stimulus type. The findings from the current experiment indicate that it may be useful to examine the effect on performance for each emotion individually, in addition to studying overall performance.

Conclusions

This experiment demonstrates that participants can reliably identify emotional vocalisations under conditions where they are not forced to respond with any of the emotion labels offered. This held true for overall performance as well as for each emotion individually. This data lends some support for research using the forced-choice paradigm. Russell's (1994) concerns regarding inflated agreement in studies using this methodology seem unfounded.

The use of the "none" label varied greatly between emotions, ranging between less than 2% for amusement and sensual pleasure stimuli to approximately 20% of the cases for anger, contentment, and sadness stimuli. These data may give some indication of which stimulus types are least likely to be perceived as belonging to an erroneous stimulus category. The addition of the "none" option only significantly affected the recognition of relief and sadness sounds. In both cases, participants were more accurate at identifying the sounds when the "none" label was not available, indicating that accuracy for these emotions may have been somewhat inflated in the forced-choice condition for these stimulus types. In sum, participants were reliably better than chance at categorising all stimulus types when a "none of the above" option was available. Thus, the addition of the "none above" option does not seem to affect overall recognition of non-verbal emotional vocalisations.

General Discussion

Positive vocal emotions

Experiment 1 demonstrated that non-verbal vocalisations of positive emotions are reliably recognisable across two language groups. These positive vocal signals were recognisable also in the context of negative emotional vocalisations (Experiment 2). This supports the notion that 'happiness' may be usefully fractionated into different, positive emotions with distinct recognisable vocalisations.

There is also evidence for some differentiation within the positive emotions: Achievement/triumph and amusement sounds were the best recognised, were rated as the most positive, and also the most aroused. In addition, they were almost never confused, either with one another or with other emotions. In contrast, there was substantial overlap between sensual pleasure and contentment in both the categorisation and the rating data in Experiments 1 and 2. Contentment was the least well recognised category of the positive emotions mainly due to contentment sounds being mistaken for sensual pleasure. Contentment sounds were also rated highly on the pleasure dimension and vice versa, suggesting that the two emotions have a high degree of common characteristics. This would seem to imply that sensual pleasure and contentment are two versions of the same emotion, perhaps belonging to a broader category, such as for example "physical enjoyment".

However, the fact that recognition of sensual pleasure sounds did not improve in the absence of the contentment category could indicate that some of the stimuli of both of these two categories were simply inadequate. This weakness could be due to the vocalisations of sensual pleasure and contentment being of relatively low intensity; previous studies have found that vocal emotions with strong emotion intensity were easier to decode than those with weak emotion intensity (Juslin & Laukka, 2001).

Negative emotions and surprise

Vocalisations of all of the "basic" emotions were recognised better than chance (Experiments 2–4). This is in line with previous work that has used non-verbal stimuli of some of these emotions (Scott et al., 1997; Schröder, 2003). The sounds rated as the most aroused of the negative emotions were anger and fear, and the categories rated as most negative were disgust and sadness. This pattern could suggest that negative emotions tend to be perceived either as strongly aroused or as very negative. This contrasts with the pattern found for the positive emotion vocalisations, where achievement/triumph and amusement sounds were rated highly for both arousal and valence, and contentment, sensual pleasure and relief were rated lower on both scales.

Surprise sounds were rated as neutral on the valence scale, and had the lowest recognition scores of the non-positive emotions. Recognition of surprise sounds improved significantly in Experiment 3, indicating that the method of stimulus selection was particularly important for the recognition of this stimulus class. The relatively low recognition rates and sensitivity to stimulus selection could indicate that surprise is not basic emotion with a reliable vocal signal. Some previous work has questioned the status of surprise as a basic emotion: In a previous study specifically investigating surprise, only a weak coherence was found between participants' spontaneous production of surprised facial expression and their cognitive, experiential and behavioural components of surprise (Reisenzein, 2000). These data were interpreted as being incompatible with a behavioural syndrome view of emotions. The behavioural syndrome view refers to the commonly held view of emotions as a set of organized patterns of behavioural, experiential, cognitive, expressive and physiological components, usually seen in an evolutionary framework, such as the basic emotion view. According to Reisenzein, the behavioural syndrome view would predict a strong probabilistic association of the different components of surprise, which is not supported by the data. Reisenzein suggests that surprise may constitute a syndrome with only weakly associated components, questioning its status as a basic emotion.

Surprise is further problematic in that several cross-cultural studies have failed to find recognition rates at higher than chance level for facial surprise stimuli, which are confused with expressions of fear (e.g., Ekman & Friesen, 1971; Ekman et al., 1969). The issue of cross-cultural recognition of surprise stimuli will be discussed in more detail in Chapter 4. There has been some discussion about whether surprise should be counted as an emotion given that it is an intrinsically valence-neutral state. Ortony and Turner (1990) suggest that "surprise is not itself an emotion, although it often plays a major role in the elicitation and intensification of emotions. When surprise is valenced, as in the case of shock, for example, the valence results from aspects of the surprising situation other than the surprise itself" (p. 317–318).

Arousal and valence

According to Russell's (1980) dimensional model, confusions and ratings of emotional stimuli should map onto the stimuli's perceived arousal and valence. This was not found to be the case, as the participants' categorisation errors and rating patterns did not consistently map onto their ratings in terms of arousal and valence. The results from the PCA identified two components underlying listeners' ratings of the stimuli. These components strongly correlated with participants' ratings of valence and arousal, and together accounted for a total of 69% of the variance in the ratings data (see Figure 2.2). This suggests that these two dimensions help form listeners internal emotional space in the sense of how emotions relate to one another, although they may not relate in a simple way to listeners' categorisations errors and ratings of the stimuli. Notably, the two factors in the PCA were highly unequal in their contribution (53% for 'valence' and 16% for 'arousal'), with the pattern suggesting a dominant role for valence. Although it can not be excluded that this could be a result of the stimulus set used, this pattern could suggest that valence is a more important aspect of emotional vocalisations than arousal. In sum, these data provide some limited support for Russell's model in that it suggests that arousal and valence underlie certain emotional judgments of how emotions relate to one another. However, as listeners' rating patterns and categorisation errors did not map simply to these dimensions, the data from this experiment cannot be said to strongly support for Russell's model.

Stimulus selection procedure

The results from Experiment 3 show that selecting the best recognised stimuli from pilot testing improves recognition. Overall recognition rates were comparable to those of previous studies using non-verbal vocal stimuli (Schröder, 2003; Scott et al., 1997). Performance for all emotions improved, although only significantly so for some emotions. Sounds of anger, fear and surprise were better recognised in Experiment 3 than in Experiment 2. Anger sounds were markedly better recognised in the current experiment as compared to previous studies (Schröder, 2003; Scott et al., 1997). It seems that recognition of anger sounds is particularly sensitive to stimulus quality, which is in line with the general pattern in Experiment 3, suggesting that recognition improves most for stimulus classes that participants have difficulty recognizing.

Forced choice and the "none" option

Experiment 4 demonstrated that adding the response option "none of the above", thereby removing the forced choice element of the task, does not significant affect overall recognition. Participants' performance remained better than chance not only overall, but also for recognition of each emotion individually. Participants used the "none" label for 10.5% of the stimuli, a somewhat higher rate than in previous studies using forced choice tasks with this added option (Frank & Stennett, 2001; Haidt & Keltner, 1999).

The addition of the "none" response option significantly reduced recognition rates for relief and sadness sounds. However, participants' performance remained significantly above chance for these stimulus types. This reduction in performance could be due to these stimuli being perceived as less similar to sounds of any other category, so that if participants perceived them not to express the correct emotion, they were less likely to categorise them as another emotion. In sum, these data are in line with previous studies using visual stimuli, lending support to research using the forced-choice paradigm. This methodology does not seem to inflate recognition rates, as the addition of the "none above" option does not reduce performance in most cases.

Summary

These experiments show that the investigations of emotional expression in the face and in speech can be applied to non-verbal expressions of emotions. "Happiness" can be fractionated into separate emotional categories with distinct vocal expressions. Participants can accurately identify non-verbal vocalisations of positive and negative emotion, and rate each emotion highest on its own scale. Valence and arousal dimensions do seem to underlie participants' emotional representations, although they do not map onto listeners' ratings and categorisation errors. These data thus cannot be said to lend support to a dimensional account of emotions. The stimulus selection method does affect the accuracy of performance, especially for emotions with lower recognition rates. Overall accuracy is not reduced when participants are offered the option of responding that none of the emotion labels match the stimulus, indicating that the forced-choice methodology is an appropriate way to test emotion recognition.

3. THE ACOUSTICS OF NON-VERBAL VOCALISATIONS OF EMOTIONS.

This chapter investigates the acoustic cues used by listeners when listening to non-verbal vocal expressions of emotions. In Experiment 5 the acoustic structure of the sounds is manipulated to determine how this affects listeners' ability to accurately identify the sounds. The results suggest a role for pitch and pitch variation in the accurate recognition of vocal expressions of emotion; a role for fine spectral structure is also implicated. The cues used when recognising non-verbal emotional vocalisations seem to differ from the cues listeners rely on when comprehending speech. In Study 6, the acoustic features of the emotional vocalisations were measured. A discriminant analysis procedure confirms that the acoustic measures provide enough discrimination between emotional categories to permit accurate automatic classification. Multiple linear regressions with participants' subjective ratings of the acoustic stimuli show that all classes of emotional ratings can be predicted by some combination of acoustic measures. In addition, all of the emotion classes are predicted by different constellations of acoustic features, with the exception of the two most highly confused emotional classes. The results of this chapter show that the acoustic cues that are used for judging the emotional contents of non-verbal vocal expressions of emotions differ from those that are used to decode speech. However, communication of emotion in both speech and non-speech sounds relies heavily on pitch cues. The perceived emotional character of the sounds can be predicted on the basis of the acoustic features of the sounds, where different sets of acoustic cues are used to predict each emotion.

Acoustic features of emotions in speech

The experiments in Chapter 2 established that naïve participants can reliably identify non-verbal vocalisations of a number of positive and negative emotions. Forced-choice and ratings tasks with emotional vocalisation stimuli cannot determine, however, what acoustic cues participants use when making these judgments. This requires a different approach, and several studies have investigated the acoustic basis of the communication of emotions in speech or speech-like utterances (e.g., Banse & Scherer, 1996; Bänziger & Scherer, 2005; Laukka, 2004; Murray & Arnott, 1993).

In an early review of emotion in speech, Murray and Arnott (1993) found that the pitch envelope was the most important parameter for differentiating between the basic emotions, whereas voice quality was the most important characteristic for distinguishing between secondary emotions such as irony, grief and tenderness. A recent study by Bänziger & Scherer (2005) investigated the role of F_0 (pitch) contour in the perception of emotional vocalisations in nonsense-speech, measuring F_0 mean, F_0 minimum, and F_0 range. They found that the F_0 level was affected by the emotional arousal of the sounds, but less by the specific emotion of the sound.

Other studies have measured a larger range of acoustic features. Banse and Scherer (1996) performed acoustic analyses measuring pitch cues, loudness, speech rate, and voiced and unvoiced average spectrum, of nonsense-speech expressing 14 emotions including hot anger, cold anger, anxiety, boredom, disgust, and happiness. They found that the emotion of the sound predicted a large proportion of the variance in most of the acoustic variables that they measured, most notably for mean energy (55% of the variance explained by emotion) and mean fundamental frequency (50% of the variance explained by emotion). Banse and Scherer also regressed the acoustic parameters onto participants' use of the emotion categories for each stimulus class in a forced-choice task. They found that for most of the emotions, this perceptual measure could be significantly predicted by some constellation of the acoustic cues (R ranged between .16 for cold anger and .27 for happiness). Three emotions could not be predicted from any combination of acoustic parameters: interest, cold anger and disgust. Banse and Scherer also performed a discriminant analysis with the measurements, in order to examine whether they would be sufficient to perform a statistical classification. A standard discriminant analysis attained 53% correct classification, whereas a more conservative jack-knifing procedure obtained a correct classification rate of 40%. The authors point out that the patterns of performance from the statistical classification analyses are very similar to those of human subjects in terms of overall performance: Human judges identified the sounds with 48% accuracy. Accuracy in the classification of individual emotions was also approximately equivalent in most cases, with the exception of a handful of emotions. In many cases, the performance of the statistical models also mirrored the kind of errors made by the human judges. The authors conclude that listeners and statistical analyses can differentiate between emotions on the basis of acoustic cues. However, they also point out that despite this similarity in the patterns of performance, there is virtually no correlation between the individual stimuli that are accurately classified by the statistical methods and those that are identified by the human judges, suggesting that the processes involved are different.

Laukka (2004) investigated the intensity of the emotional expressions as well as emotional content. He measured 20 voice cues of speech expressing five emotions, each at two levels of intensity (weak and strong). The results showed that emotion and intensity both had an effect on most of the voice cues measured. Laukka also performed a series of multiple regressions with nine of the acoustic measurements as independent variables, and the participants' ratings of the sounds on emotional rating scales as dependent variables. All of the ratings could be significantly predicted by using the acoustic cues of the sounds, with a unique constellation of cues predicting each emotion. Laukka concludes that "listeners use emotion-specific patterns of cues to decode emotion portrayals" (p. 38). In additional analyses, the study found that the intensity of the emotional expressions could also be predicted from the acoustic features of the sounds.

Aims of this chapter

Research into the acoustic features of emotion in speech has generally shown a dominant role for pitch and pitch variation in the expression of emotions in verbal sequences, as well as intensity, and spectral qualities such as tension. However, little is known about what acoustic features of non-verbal emotional vocalisations are important in order for the listener to decode the intended emotional message.

The aim of Experiment 5 is to experimentally investigate how acoustic factors affect the perception of non-verbal emotional expressions. The acoustic structure of the emotional noises is manipulated, in order to selectively alter different acoustic characteristics. The effect of these different manipulations on participants' ability to identify the emotional sounds is then evaluated using a forced-choice task. Unlike previous work (e.g., Schröder, 2003), this method allows an investigation of the relationship between acoustic structure and the *recognition* of emotions.

Study 6 aims to map out the acoustic features of the (original) non-verbal emotional sounds in terms of pitch cues, spectral cues and amplitude envelope information. First, these measurements are used in discriminant analyses in order to examine whether the information provides sufficient detail to discriminate between the different emotion categories. The acoustic measurements are then used to test whether the acoustic features of the emotional sounds can predict participants' perception of the sounds, using the rating data from Experiment 2.

Experiment 5 — Recognition of emotions in distorted sounds

Human vocalisations contain a number of acoustic variables, including duration, intensity, amplitude modulation, pitch and pitch variation, as well as spectral structure (timbre), which can be characterised as fine or broad spectral structure. Different kinds of information are carried by fine and broad spectral structure: speech intelligibility is possible with only the broad spectral information preserved (Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995), while polyphonic music is well recognised from fine spectral structure alone (Smith, Delgutte, & Oxenham, 2002). In the current experiment, techniques were used that selectively removed or altered several of these characteristics, to investigate how this affects listeners' classification of these sounds.

Three different transformations were selected to manipulate aspects of the acoustic structure of the sounds. The first transformation was one-channel noise-vocoding (Shannon et al., 1995). This manipulation removes all the acoustic information except the amplitude envelope of the sounds — there is no pitch and no spectral structure (fine or broad), but the rhythm and duration of the original sound is preserved. There is a linear relationship between noise-vocoding and speech intelligibility: the more channels, the better the speech can be understood (Faulkner, Rosen, & Wilkinson, 2001), but one-channel noise-vocoding conveys only the rhythm of speech and cannot be understood (Rosen, 1992).

The second transformation was six-channel noise-vocoding, a technique which simulates a cochlear implant with six electrodes (Shannon et al., 1995). This technique removes most of the pitch and all the fine spectral structure, leaving the duration, the rhythm and the broad spectral structure. This transformation was chosen because after a short training period, six-channel noise-vocoded speech becomes easily intelligible — that is, the semantic/linguistic content of the speech becomes decodable (Scott, Blank, Rosen, & Wise, 2000). This is the case despite the fact that six-channel vocoded speech lacks fine spectral structure (e.g., speaker identity information is lacking) and has such a weak pitch that the speech has little or no "melody". If the same acoustic cues are used to recognise emotional expression as to decode intelligible speech, the emotional sounds in this condition would be expected to be well recognised.

The third transformation was spectral rotation (Blesser, 1972). This technique can be considered analogous to inversion of facial stimuli in that the same physical information is present, but the global configuration is radically altered. Spectral rotation preserves amplitude envelope and duration information, pitch and pitch variation, while distorting both fine and broad spectral

Tab. 3.1: Acoustic information of normal and acoustically manipulated stimuli. Y indicates a preserved feature, and N indicates removed a feature. Speech Intel = speech intelligibility.

Manipulation	Pitch & Intonation	Spectral Detail	Rhythm	Speech Intel
Original	Y	Y	Y	Y
1-channel noise-vocoded	N	N	Y	N
6-channel noise-vocoded	Weak	Some	Y	Y
Spectrally rotated	Y	Y, but on wrong place on frequency dimension	Y	X

information. Thus, while the intonation profile of spectrally rotated speech is preserved, the speech cannot be understood due to the gross distortion of the spectral structure (Blesser, 1972). If the recognition of the emotional stimuli is strongly linked to the sense of pitch and pitch variation, the emotional sounds in this condition would be expected to be well recognised. The fourth manipulation was not a transformation per se but filtered versions of the non-verbal vocal expressions of emotion used in the previous experiments, and in which the amplitude envelope, duration, spectral structure (fine and broad) and pitch are unchanged. Table 3.1 summarizes the four different conditions.

These conditions are by no means exhaustive as possible manipulations of the sounds; for example, the amplitude envelope or durations of the sounds are not altered in any of these manipulations. Rather, these conditions were selected because of the established impact they have on speech intelligibility, in order to determine the similarities or differences that exist between the acoustic factors important for speech intelligibility and those used for understanding non-verbal emotional expressions. These stimuli (normal and transformed) were presented to naïve participants in order to determine listeners' ability to recognise emotion in these conditions. The results can tell us what cues are important for recognising non-verbal expressions of emotions, and how these might vary between recognition of speech intelligibility and non-verbal stimuli.

Three specific hypotheses were tested: It was predicted that participants would perform better with the original stimuli than with the manipulated stimuli, as the former contain more acoustic information than the latter. This experiment also examined whether six-channel noise-vocoded stimuli would be better recognised than the spectrally rotated stimuli sounds. As mentioned, research has established that spectrally rotated speech is unintelligible, whereas six-channel noise-vocoded speech can be understood with minimal training. It may be that spectral rotation degrades the acoustic information necessary for listeners to decode the signal, affecting the information necessary to judge emotional cues as well as speech content. The third hypothesis predicted that six-channel noise-vocoded stimuli would be better recognised than one-channel noise-vocoded sounds. Six-channel noise-vocoded stimuli are richer in acoustic information, including pitch cues which are thought to be important in emotional communication in the voice (Murray & Arnott, 1993). It is also well established that speech intelligibility is much better with six- than one-channel noise-vocoded stimuli (Faulkner et al., 2001).

Methods

Stimuli

Fifty stimuli (5 randomly selected exemplars per emotion) from Experiment 2 were included. All stimuli were low-pass filtered at 4 kHz, allowing the spectral rotation to be performed. Low pass filtering at this threshold does not perceptibly alter the stimuli. Copies of the stimuli were acoustically manipulated in three different ways in order to remove different acoustic types of information (see Table 3.1), resulting in a stimulus set of 200 sounds.

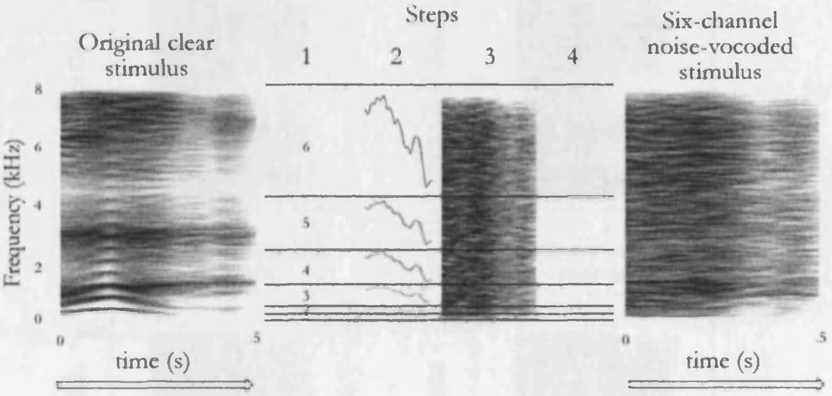

Fig. 3.1: Illustration of the processing steps involved in noise-vocoding with six channels. Step 1 involves filtering the signal into six frequency ranges, from which the amplitude envelopes are extracted (Step 2). In Step 3, the noise in each frequency range is modulated using these amplitude envelopes, which are then combined (Step 4), producing the noise-vocoded stimulus. Adapted from Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan (2005).

Two sets of stimuli were noise-vocoded using PRAAT (Boersma & Weenink, 2005). In noise-vocoding, the signal is band-pass filtered into a number of frequency bands (channels). The amplitude envelope of each band is extracted and applied to band-pass filtered noise in the same frequency. Then the separate bands of modulated noise are recombined into the final stimulus.

One set of stimuli was noise-vocoded using one channel and another set with six channels. This process is illustrated in Figure 3.1.

A third set was spectrally rotated around a 2 Hz mid-point, using a version of the technique described by Blesser (1972). In this manipulation the spectral detail of the sound is inverted. This is done by high-pass filtering the signal, amplitude-modulating it with a sinusoid around 4 Hz, and then low-pass filtering the signal. This manipulation preserves the amplitude envelope and the overall spectral and pitch variation, although the pitch sensation is weaker (Blesser, 1972).

Figure 3.2 shows spectrograms and oscillograms for one disgust sound in the three manipulations and the original (low-pass filtered) stimulus. The amplitude envelope and duration of the sound remain unchanged by these manipulations. There were 5 stimuli per condition expressing achievement/triumph, amusement, anger, contentment, disgust, fear, sensual pleasure, relief, sadness, and surprise. Examples of the stimuli are available in the additional materials.

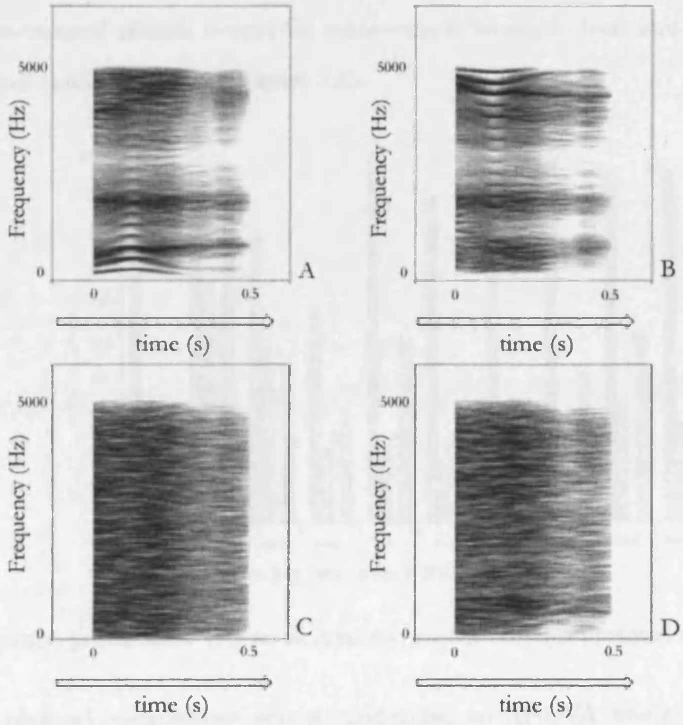

Fig. 3.2: Spectrograms of a disgust stimulus in various manipulations: original (A), spectrally rotated (B), one-channel noise-vocoded (C), and six-channel noise-vocoded(D).

Participants & Procedure

Seventeen participants (8 male; mean age 25 years), participated in a 10-way forced-choice task where the order of the stimuli was randomised for each participant across emotional and acoustic manipulation conditions (i.e., the presentations were not blocked). The procedures were the same as the categorisation task in Experiment 2. The participants had not taken part in any previous study of vocal emotions.

Results

Chi-square analyses of correct and incorrect responses were used to determine whether participants were categorizing the stimuli at levels reliably above chance (see Appendix E). The results showed that the original and rotated stimuli were all categorised at rates above chance. Most of the six-channel stimuli were categorised above chance (all apart from achievement/triumph sounds) and most of the one-channel stimuli except for achievement/triumph, fear, and sensual pleasure, were recognised better than chance (see Figure 3.3).

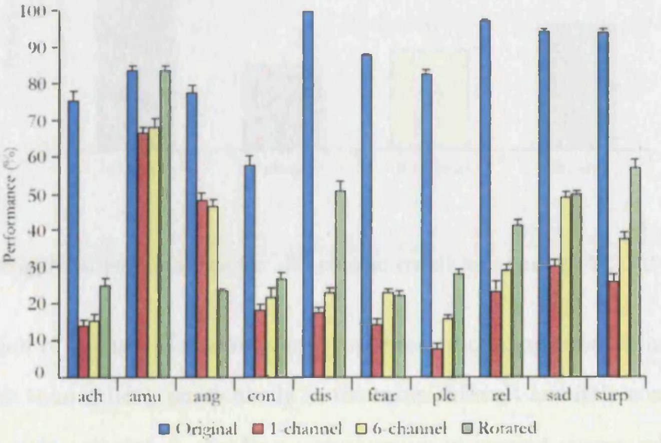

Fig. 3.3: Recognition performance (%) for normal and acoustically manipulated stimuli for each emotion.

To enable planned comparisons across conditions, an ANOVA was also performed, treating the correct responses as scores (out of five for each participant for each stimulus type). This revealed a significant main effect of emotion ($F_{(9,144)}$ =18.2, p 0.0001), a significant main effect of manipulation ($F_{(3,48)}$ =273.2, p 0.0001) and a significant emotion by manipulation interaction ($F_{(27,432)}$ =6.4, p 0.0001; see Figure 3.3).

Three planned comparisons were carried out. The first contrasted the original stimuli with all of the manipulated conditions, to test the prediction that the original stimuli would be the more accurately categorised. This contrast was significant($F_{(1)}$ =686.7, p 0.0001) (mean score for original stimuli = 3.80, aggregate mean score for the manipulated conditions = 1.39; see Figure 3.4). The second comparison contrasted the rotated and six-channel noise-vocoded stimuli, since in the case of speech the six-channel stimuli would be far better comprehended than the rotated stimuli. This contrast revealed a significant difference ($F_{(1)}$ =10.2, p 0.005), but the difference was in the opposite direction (mean recognition for rotated stimuli = 1.69, mean recognition for the six-channel = 1.34, see Figure 3.4). The third contrast was the one-channel versus six-channel, since if the stimuli were speech, the six-channel stimuli would be much better recognised than one-channel stimuli (Faulkner, Rosen, & Wilkinson, 2001). This was only a weak trend ($F_{(1)}$ =3.0, p = 0.09; mean recognition for one-channel stimuli = 1.14, see Figure 3.4).

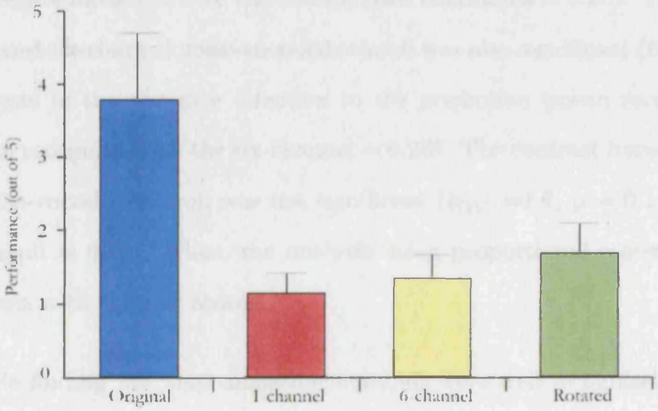

Fig. 3.4: Average recognition performance for all acoustic conditions (out of 5). All emotions collapsed.

A visual inspection of the use of the different response options indicated that some of the labels were used more often than others, particularly in the more difficult acoustic conditions. Therefore, proportional scores were calculated, yielding a proportion of correct scores relative to the use of each response label.

The ANOVA and planned comparisons were repeated using the proportional accuracy scores. This analysis showed a significant main effect of emotion ($F_{(9,144)}$ =8.9, p 0.0001), a significant main effect of manipulation ($F_{(3,48)}$ =166.6, p 0.0001), and a significant emotion by manipulation interaction ($F_{(27,432)}$ =2.8, p 0.0001; see Figure 3.5).

Three planned comparisons were carried out. The contrast between the original stimuli and

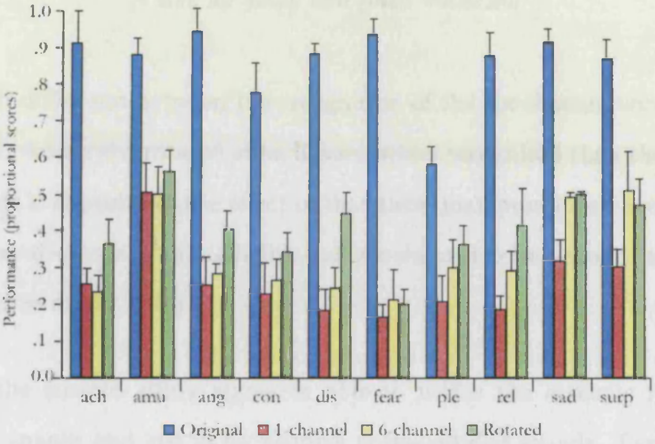

Fig. 3.5: Proportional scores of recognition performance relative to label use in each acoustic condition.

all of the manipulated conditions was significant ($t_{(16)}$ =19.7, p < 0.0001; mean score for original stimuli = 0.76, aggregate mean score for the manipulated conditions = 0.27). The contrast between the rotated sounds and six-channel noise-vocoded stimuli was also significant ($t_{(16)}$ =2.8, p < 0.05), but the difference was in the opposite direction to the prediction (mean recognition for rotated stimuli = 0.33, mean recognition for the six-channel = 0.26). The contrast between the one-channel and six-channel noise-vocoded stimuli was not significant ($t_{(16)}$ =1.7, p = 0.11; mean recognition for one-channel stimuli = 0.22). Thus, the analysis using proportional scores yielded equivalent results to the analysis with the raw scores.

One final notable finding was that amusement sounds were well recognised across all acoustic conditions. This likely contributes to the significant emotion x manipulation interaction effect and the main effect of emotion. It appears that none of these acoustic manipulations were sufficient to disrupt the perception of amusement (see Figure 3.5), as even a one-channel noise-vocoded "laugh" is still well recognised as an amusement sound. Since a one-channel noise-vocoded sound contains only amplitude envelope variation, and no pitch or spectral detail, this indicates that properties of the amplitude envelope alone are sufficient to convey amusement.

Discussion

There were clear effects of acoustic manipulations on the recognition of emotion in non-verbal vocalisations. While all the manipulated sounds were less well recognised than the original stimuli, there were differences within these which indicate some of the acoustic cues that the participants were using to identify emotions.

A role for pitch and pitch variation

There was an overall difference between the recognition of the six-channel vocoded and the spectrally rotated stimuli, where the rotated stimuli were better recognised than the six-channel noise-vocoded stimuli. This is opposite to the effect of that these manipulations have on the recognition of speech, where rotated speech is unintelligible and six-channel noise-vocoded speech is intelligible (Blesser, 1972; Shannon et al., 1995).

Participants in the current study appeared able to utilise the acoustic information that is preserved in rotated sounds and lost in six-channel noise-vocoded sounds. Two related candidate acoustic cues are pitch and pitch variation: pitch and pitch movement are preserved in rotated sounds and very weak in the six-channel noise-vocoded sounds. Previous studies with emotional speech (Murray & Arnott, 1993) have indicated that such intonation cues are important for the perception of emotion (not intelligibility) in speech. The current study is the first demonstration that this may also be the case for non-verbal vocal expressions of emotion. This suggests that the acoustic correlates of the identity of emotion in the voice may be very different from the acoustic features important for understanding the meaning of speech, where broad spectral variation is among the most important factors.

Spectral cues

There were minimal differences between the one-channel and six-channel noise-vocoded sounds. This suggests that there may be a smaller contribution of the broad spectral structure of the sounds to their emotional categorisation. It may be that the contribution of spectral properties is more relevant at the level of fine spectral detail (not present in either noise-vocoded conditions, nor preserved in rotated speech). Consistent with this, there were clear differences between the original sounds and their rotated equivalents; the latter maintain the pitch and pitch variation (as well as duration and amplitude variation) of their untransformed equivalents, but invert their spectral detail (broad and fine). This suggests that fine spectral detail is important for the recognition of emotion in non-verbal utterances (and probably verbal utterances), while the broad spectral structure (i.e., formants) is crucial for the intelligibility of speech but not emotion. Certainly fine spectral detail captures properties of voice quality, such as tension, which contribute to the emotional quality of vocalisations (Murray & Arnott, 1993).

The relative contributions of different cues

Further studies will be needed to assess the relative contributions of pitch, pitch variation and fine spectral detail in the recognition of emotion in the voice. It is difficult to preserve fine spectral structure of sounds while removing pitch and pitch variation cues (as fine spectral structure is one cue which contributes to pitch perception); the converse manipulation is also hard to make. However, we can see that *without* preserved fine spectral detail, pitch and pitch variation helps recognition, as rotated stimuli are better recognised than one- and six-channel noise-vocoded stimuli. Performance for the rotated sounds is less than half of the recognition rate for the untransformed original stimuli (see Figure 3.4), which implies that spectral cues do serve an important function. So although there is a role for pitch and pitch variation in the recognition of emotional stimuli, there is also clearly an important, and possibly even dominant, role for fine spectral structure.

Previous studies have identified pitch and aspects of pitch variation as important cues for emotion in speech (Murray & Arnott, 1993; Bänziger & Scherer, 2005). These findings appear to confirm a role for pitch for perception of emotion in non-verbal vocalisations. They also indicate that the fine spectral structure of sounds is important in determining their emotional identity. Fine spectral structure is associated with pitch, and has been shown to be important in the perception of music (Smith et al., 2002). Further studies will elucidate the precise contributions and interaction between fine spectral detail and pitch information.

Specific emotions

There were also some clear interactions between acoustic condition and emotion condition. Most noticeably, amusement was well recognised across all conditions (original = 98%, rotated = 83%, one- and six-channel = 64%). Since all the acoustic manipulations preserved amplitude variation and duration, this suggests that amusement is carried predominantly by these cues. As in previous experiments, contentment was the emotion that was least well recognised across all conditions.

Speech and non-verbal expressions

The stimuli in this study were deliberately selected to be non-verbal, but are there any similarities between the acoustic factors used to identify the emotional information carried in these stimuli and those used for the identification of lexical or emotional information in speech? The difference from lexical decoding is striking: the perception of meaning in speech is highly dependent on the patterns of gross spectral structure (e.g., formants) and how these change over time. Unless a language uses lexical tones, or voice quality contrastively, pitch and fine spectral structure are not directly relevant for lexical identity. In this study, the behavioural measures indicate that pitch and fine spectral structure are key components in the processing of emotional vocalisations, while gross spectral structure appears to have only marginal effects (perhaps in the recognition of angry sounds); see Figure 3.3. However, there seems to be more similarity between the detection and recognition of emotion expressed in speech and the recognition of these non-verbal stimuli. Emotion in speech has been linked to pitch variation (Bänziger & Scherer, 2005), as is the case in this set of non-verbal stimuli. Likewise, a role for voice quality has been ascribed to the perception of emotion in speech (Murray & Arnott, 1992), as is the case in these non-verbal stimuli. In conclusion, pitch and fine spectral detail are important cues for emotion in speech and non-verbal expressions, while broad spectral structure is important for establishing lexical information in speech.

Study 6 — Acoustic measurements

The results of Experiment 5 indicate a role for pitch variation and fine spectral structure in the identification of non-verbal emotional vocalisations. They also suggest that there may be differences between the types of acoustic cues that are used for identifying different emotions. This experiment provides an acoustic analysis of the non-verbal emotional vocalisations, aiming to describe the ways in which sounds from different categories vary on a number of acoustic features. This approach has been used in the study of both facial expressions of emotions (Calder et al., 2001) and emotions in speech (Laukka, 2004). Calder et al. demonstrated that a principal component analysis (PCA) of the pixel intensities of facial expressions produced a set of principal components (PCs) that could both categorise individual emotions effectively and capture their rated values on arousal and valence dimensions. In an investigation of emotions in speech, Laukka (2004) measured 20 voice cues of speech with weak and strong intensity in five emotions. He found that participants' emotional ratings of the sounds could be reliably predicted by the acoustic cues.

The current study used measured aspects of intensity, amplitude envelope, duration, spectral centre of gravity, pitch, and pitch variation of the stimuli from each emotional category. To determine whether these measurements capture sufficient detail of the sounds to classify them, discriminant analyses procedures were employed. The acoustic measurements were then used in a series of multiple linear regressions to predict the emotional ratings of the stimuli in Experiment 2. The aim was to identify which patterns of acoustic variation were associated with the perceived emotional contents of the sounds. To reiterate, two procedures were used to capture psychophysical properties of both input and output of emotional vocalisations: discriminant analysis to assess whether acoustic qualities can sort stimuli on the basis of the speaker's intent, and linear regressions using rating data to assess whether acoustic qualities can predict the listener's judgments.

Methods

Stimulus Measurements

We took measurements of acoustic parameters from 165 sound files (16 – 17 of each emotion) using the PRAAT program (Boersma & Weenink, 2005). All sounds were downsampled to 44.1 kHz and scaled to have the same peak amplitude (0.291 Pa; the mean peak amplitude of the original recordings) prior to the analysis. This was necessary since the wide dynamic range of the original stimuli meant that the recording levels were different across the different emotional conditions. This still permits the analysis of intensity variation, which is computed across the whole waveform.

In the amplitude domain, standard deviation, duration, and mean intensity (dB SPL scale) were obtained from the waveform. The number of amplitude onsets per sound file were counted, using an algorithm that detects local rises in the smoothed amplitude envelope (Cummins & Port, 1998). This gives an estimate of the number of 'syllables' in a vocalisation. To this end, the signal was first band-pass filtered (Hanning filter centred at 2.2 kHz with a bandwidth of 3.6 kHz), full-wave rectified, and smoothed (Hanning low-pass filter with an 8-Hz cutoff) before obtaining the first derivative of the smoothed envelope. Onsets were then defined as points in time at which (a) a set threshold in the amplitude envelope was exceeded and (b) the derivative curve had a positive value.

Pitch measurements were based on a derived curve representing change in fundamental frequency as a function of time (using a 75 – 1000 Hz analysis window and the autocorrelation method described in Boersma, 1993). From this, minimum, maximum, mean, and standard deviation were obtained. Global pitch movement was defined as the difference between the mean pitch values of the first and last 20% of the sound file. However, pitch movement was dropped as a variable as this measurement could not be calculated for any of the relief or surprise stimuli (due to their brevity) or for half of the anger stimuli (due to their weak pitch). Finally, we computed the spectral centre of gravity and the standard deviation of the spectrum on the basis of fast Fourier transformations.

Statistical Procedure

Discriminant analysis Discriminant Analyses was performed in order to examine whether the acoustic measurements from the PRAAT analysis provided sufficient information to distinguish between emotional categories. The independent variables of the models were the acoustic measurements of the non-verbal vocalisations of the 165 stimuli, and the dependent variable was the emotion category selected by the discriminant analyses. Discriminant analysis identifies a set of functions that minimises within-category variability and maximises between-category variability. These functions are used by the model to predict the category membership of each of the stimuli in the set.

As the standard method of discriminant analysis tends to overestimate the accuracy of the model (Calder, Burton, et al., 2001), the more conservative jack-knifing method was also used. With jack-knifing, one analysis is performed for each stimulus. In each analysis, the category membership of all but one of the stimuli is known and the model predicts the membership of the uncategorised stimulus. This prediction is made based on the discriminant functions derived from the other, categorised, stimuli. In this study, 165 analyses were carried out, one for each stimulus. The performance of the model is measured by the percentage of categorisations made by the model that are correct, which can be assessed both overall and for each emotion category separately. If the accuracy of the model's classifications is high, this indicates that the independent variables are sufficient to distinguish between the categories. In this case this would indicate that the acoustic measurements of the sounds from the PRAAT analysis provide sufficient detail to distinguish between the different emotion classes.

Multiple regressions In order to determine which acoustical characteristics underlie judgments of each of the emotions, stepwise multiple regressions were performed for each of the emotional rating scales from Experiment 2 (achievement/triumph, amusement, anger, contentment, disgust, fear, sensual pleasure, relief, sadness and surprise). This analysis only included the acoustic measurements of the 100 stimuli rated in Experiment 2, as this rating data was required as dependent variables in the regression. The independent variables were the acoustic measurements. These analyses show whether any constellation of acoustic measurements from the PRAAT analysis significantly predict participants' ratings on each of these emotional rating scales. Multiple regressions were also carried out with the participants' ratings on the arousal and valence scales in Experiment 2, in order to determine whether the acoustic measurements of the sounds accurately predict these perceived qualities.

Results

Acoustic analysis and discriminant analysis

The results of the acoustic analysis are displayed in Appendix F. The results of the standard discriminant analysis (above) and the jack-knifing analysis (below) are shown in Table 3.2. The overall accuracy of the analyses was 56.4% for the standard discriminant analysis and 50.3% for the jack-knifing analysis.

Chi-square analyses were performed to test whether the models' overall performance was better than would be expected by chance (10% accuracy). The results indicated that the acoustic measurements provide sufficient information to discriminate successfully between stimuli from different emotional categories for both types of discriminant analyses ($\chi_{(9)}$ = 814.3 for the standard analysis, 595.3 for the jack-knifing analysis, both $p < 0.0001$). In the case of the standard discriminant analysis, performance was lowest for contentment (23.5%) and fear (31.3%) and highest for amusement (76.5%) and surprise (81.3%). In the jack-knifing analysis, performance was lowest for contentment (29.4%) and fear (25.0%) and highest for achievement/triumph (70.6%).

Sets of chi-square analyses were also performed to examine whether the models performed significantly better than chance in classifying stimuli from each of the emotional categories. For the standard discriminant analysis, the model performed significantly above chance for all classes, except for contentment sounds ($\chi_{(9)}$ = 73.0 for achievement/triumph, 87.1 for amusement, 44.0 for anger, 25.3 for disgust, 26.5 for fear, 49.5 for sensual pleasure, 61.2 for relief, 44.0 for sadness, and 91.5 for surprise, all $p < 0.05$, Bonferroni corrected for 10 comparisons). For the jack-knifing analysis, the model performed significantly above chance for all classes, except for contentment

Tab. 3.2: Results of standard (above) and jack-knifing (below) discriminant analysis for classification of non-verbal emotional vocalisations from acoustic analysis. All results in %, correct classifications in bold. All horizontal rows add to 100. Ach = Achievement/Triumph, Amu = Amusement, Ang = Anger, Cont = Contentment, Dis = Disgust, Ple = Pleasure, Rel = Relief, Sad = Sadness, Surp = Surprise.

Stimulus type	Classification									
	Standard discriminant analysis									
	Ach	Amu	Ang	Con	Dis	Fear	Ple	Rel	Sad	Sur
Achievement	**70.6**	0.0	11.8	0.0	5.9	11.8	0.0	0.0	0.0	0.0
Amusement	0.0	**76.5**	0.0	11.8	0.0	0.0	0.0	0.0	11.8	0.0
Anger	6.3	0.0	**56.3**	0.0	18.8	12.5	6.3	0.0	0.0	0.0
Contentment	0.0	11.8	5.9	**23.5**	0.0	0.0	29.4	17.6	11.8	0.0
Disgust	6.3	6.3	18.8	0.0	**43.8**	0.0	6.3	12.5	6.3	0.0
Fear	31.3	0.0	25.0	0.0	6.3	**31.3**	0.0	0.0	0.0	6.3
Pleasure	0.0	0.0	0.0	11.8	0.0	0.0	**58.8**	11.8	5.9	11.8
Relief	0.0	0.0	5.9	17.6	0.0	0.0	5.9	**64.7**	0.0	5.9
Sadness	0.0	18.8	0.0	6.3	0.0	0.0	6.3	0.0	**56.3**	12.5
Surprise	0.0	0.0	0.0	0.0	6.3	6.3	0.0	6.3	0.0	**81.3**
	Jack-knifing									
Achievement	**70.6**	0.0	11.8	0.0	5.9	11.8	0.0	0.0	0.0	0.0
Amusement	0.0	**58.8**	0.0	11.8	0.0	0.0	0.0	11.8	17.6	0.0
Anger	6.3	0.0	**56.3**	0.0	18.8	12.5	6.3	0.0	0.0	0.0
Contentment	0.0	11.8	0.0	**29.4**	0.0	0.0	29.4	17.6	11.8	0.0
Disgust	6.3	6.3	18.8	0.0	**43.8**	0.0	6.3	12.5	6.3	0.0
Fear	31.3	0.0	25.0	0.0	12.5	**25.0**	0.0	0.0	0.0	6.3
Pleasure	0.0	0.0	0.0	11.8	0.0	0.0	**58.8**	11.8	5.9	11.8
Relief	0.0	0.0	5.9	17.6	0.0	0.0	5.9	**58.8**	5.9	5.9
Sadness	0.0	18.8	0.0	6.3	0.0	0.0	12.5	0.0	**50.0**	12.5
Surprise	0.0	0.0	0.0	0.0	18.8	12.5	0.0	18.8	0.0	**50.0**

and fear sounds ($\chi_{(9)}$ = 73.0 for achievement/triumph, 51.8 for amusement, 44.0 for anger, 25.3 for disgust, 49.5 for sensual pleasure, 49.5 for relief, 35.3 for sadness, and 37.8 for surprise, all $p < 0.05$, Bonferroni corrected for 10 comparisons).

Common confusions in the standard discriminant analysis were anger sounds categorised as disgust, contentment sounds categorised as sensual pleasure and relief, disgust sounds categorised as anger, fear sounds categorised as achievement and anger, relief sounds categorised as contentment, and sadness sounds categorised as amusement. The confusions in the jack-knifing analysis were similar, except that amusement sounds were also categorised as sadness, and surprise sounds were categorised as disgust and relief. The overall pattern of results was consistent with that of human participants in Experiment 2 (see Table 2.3), although the human performance was somewhat higher. Nevertheless, the discriminant analyses clearly demonstrate that the acoustic measurements provide sufficient information to categorise the emotional sounds accurately.

Multiple regressions

The regression analyses were significant for all of the emotional scales (see Table 3.3). This indicates that the participants' ratings on each of the emotional scales in Experiment 2 could be reliably predicted from the acoustic measurements of the sounds. The variance explained by the acoustic factors ranged from 16% for the achievement/triumph ratings to 37% for the ratings on the fear scale. Clearly, much of the variance in the emotion ratings is due to other factors than the acoustic measurements captured in the PRAAT analysis. Still, these measurements predict a significant portion of the participants' ratings on each of these scales.

Tab. 3.3: Results of step-wise multiple regression analyses for each of the ratings scales from Experiment 2. Ampons = Amplitude onsets, Ampstd = Amplitude standard deviation, Dur = Duration, Int = Intensity, Pstd = Pitch standard deviation, Pmean = Pitch mean, Spcog = Spectral centre of gravity, Spstd = Spectral standard deviation.

Emotion	Acoustic features	Adjusted R^2	F	Df	P-level
Achievement	Ampstd, spstd	.16	10.5	2,96	0.000
Amusement	Ampons, amstd, spstd	.20	9.3	3,95	0.000
Anger	Spcog, spstd, pmean	.27	12.9	3,95	0.000
Contentment	Spcog, dur, int, pstd	.32	12.5	4,94	0.000
Disgust	Spcog, dur, ptstd, spstd	.26	9.6	4,94	0.000
Fear	Pmin, spcog, spstd, pmean	.37	15.6	4,94	0.000
Pleasure	Spcog,dur,int,spstd	.33	10.1	4,94	0.000
Relief	Spstd, spcog, int	.26	12.6	3,95	0.000
Sadness	Int, spstd	.19	12.2	2,96	0.000
Surprise	Pmean, dur	.20	12.9	2,96	0.000
Arousal	Spcog, spstd, ampstd, ampons	.55	25.5	4,94	0.000
Valence	Ampstd, dur	.08	5.0	2,96	0.008

As can be seen in Table 3.3, each emotional rating scale is predicted by a distinct constellation of factors. Indeed, no two scales have the same pattern of factors, with the exception of contentment and sensual pleasure. The acoustic factors can be usefully divided into envelope cues, pitch cues and spectral cues. Figure 3.6 is a schematic illustration of types of cues that predict the ratings on each of the emotional scales. As can be seen in the Venn diagram in Figure 3.6, none of the emotional scales are judged on the basis of only one class of acoustic information: Fear and anger ratings are based on spectral and pitch information, surprise on the basis of pitch and envelope cues, disgust on the basis of all three classes of acoustic information, and sadness and all of the positive rating scales on the basis of spectral and envelope cues.

Multiple regressions were also carried out for the participants' ratings on the arousal and valence scales in Experiment 2 (see Table 3.3). The model using the valence ratings was significant but predicted only 8% of the variance of these ratings. The model using the arousal ratings was

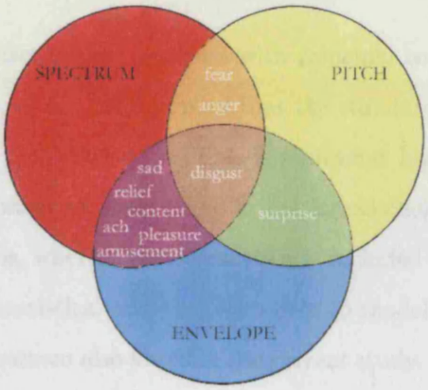

Fig. 3.6: Venn diagram of classes of acoustic information used to predict participants ratings for various emotional scales.

significant and explained 55% of the variance. This model included spectral centre of gravity, spectral standard deviation, amplitude standard deviation, and number of amplitude onsets.

Discussion

The results of this study suggest that, as with facial stimuli and emotion in speech, there is a mapping between stimulus characteristics and emotional category in non-verbal vocalisations. In addition, different emotions are perceived on the basis of different acoustic information.

Discriminant analyses

The discriminant analysis and the jack-knifing analysis were accurate at categorising the emotional vocalisations (56.4% and 50.3% correct, respectively). This demonstrates that the measurements of the acoustic analysis provided sufficient information to successfully discriminate between stimuli from different emotional categories. The statistical models performed significantly better than chance not only in terms of overall performance but also in classifying stimuli from each of the emotional categories. Exceptions to this were the contentment sounds which were not classified at rates that exceeded chance in either analysis, and fear sounds, which were not classified above chance in the jack-knifing analysis. The pattern of confusions broadly mirrored those found in Experiment 2, although the human performance was somewhat higher. In sum, the discriminant analyses clearly demonstrate that the acoustic measurements from the PRAAT analysis provide sufficient information to categorise the emotional sounds accurately.

In a study which used discriminant analyses with principal components of facial expressions of emotions, Calder, Burton, et al. (2001) found that the standard model had an accuracy score of 67% and the jack-knifing procedure 53%. This is somewhat higher than in the current study, which may be due to the number of alternatives in the forced-choice task: Calder et al. used six categories of facial expressions, whereas the current study included ten emotion categories. Calder et al. (2001) point out the statistical analyses were able to model human participants' errors as well as their performance, a pattern also found in the current study. Banse and Scherer (1996) used both discriminant analysis and jack-knifing with nonsense-speech expressing 14 different emotions. The standard discriminant analysis classified the sounds correctly in 53% of the cases, whereas the jack-knifing procedure reached 40% accuracy, both close to the human performance at 48%. Again, the performance of the models also mirrored the kind of errors that were made by the human judges. It seems thus that for both facial and vocal expressions of emotions, it is possible to classify emotional expressions modeling human performance, on the basis of basic perceptual cues.

Multiple regressions

The participants' ratings on each of the emotion scales could be predicted from the acoustic measurements of the sounds, with a particular constellation of acoustic cues for each emotional scale. Exceptions were contentment and sensual pleasure, which were predicted by the same constellation of acoustic cues. None of the emotional scales were judged on the basis of only one class of acoustic information: Fear and anger were rated based on spectral and pitch information, surprise was rated from pitch and envelope cues, and disgust from all three classes of acoustic information. Sadness and all of the positive rating scales were rated on the basis of spectral and envelope cues. The variance explained by the acoustic factors ranged from 16% for the achievement/triumph ratings to 37% for the ratings on the fear scale. The variance in the emotion ratings is likely due to other factors than the acoustic measurements captured in the PRAAT, including cues such as voice quality.

These findings are in line with previous findings from a study with emotional speech. Laukka (2004) regressed nine acoustic measures onto participants' mean ratings of speech sounds on emotional rating scales. A unique constellation of cues predicted each emotion, with R^2 ranging between 45% for fear and happiness to 58% for anger. The amount of variance accounted for by the acoustic cues in Laukka's study was higher than in the current study, which may be due to differences between verbal and non-verbal vocalisations of emotions. Speech segments are longer and allow for additional measurements such as speech rate and proportion of pauses.

In a study using nonsense-speech, Banse and Scherer (1996) regressed acoustic parameters onto participants' use of the emotion categories for each stimulus class in a forced-choice task. They found that for all of the emotions except three, the participants' categorisations could be significantly predicted from the acoustic cues (R ranged between .16 for cold anger and .27 for happiness). Banse and Scherer also did a reverse set of regressions, attempting to predict acoustic features from the emotion of the sounds. They found that the emotion of the sounds predicted a large proportion of the variance for most of the acoustic variables that they measured.

Arousal and valence

More than half of the variance in the participants' arousal ratings could be predicted from the acoustic features of the sounds. The acoustic cues were the spectral centre of gravity, spectral standard deviation, amplitude standard deviation and amplitude onsets. This provides support for an acoustic "arousal" dimension, consistent with previous claims in the literature that vocal communication has a dominant role for non-specific arousal cues (Bachorowski, 1999; Bänziger & Scherer, 2005). Notably, the stronger dimension in the PCA analysis of ratings data (in Experiment 2) was valence (explaining 53.3% of the variance). In the current analysis, however, there was not a strong relationship between the acoustic cues and the valence ratings. This concurs with a previous study using emotional sounds, in which Laukka (2004) regressed acoustic cues onto listeners' dimensional ratings of the sounds. He found that the acoustic cues predicted markedly less variance for the participants' valence ratings compared to all of the other rating scales. This could lend some support to Laukka's suggestion that valence may reflect a categorical difference between positive and negative sounds rather than a true dimension.

General Discussion

Acoustic features

Responses to the acoustically manipulated stimuli from Experiment 5 indicated a dominant role for pitch and pitch variation in emotion recognition, which was indicated by higher accuracy for the spectrally rotated stimuli compared to the noise-vocoded sounds. The results from Experiment 5 also point towards a role for fine spectral structure, in the difference between the spectrally rotated and the original sounds. A smaller contribution comes from broad spectral detail, shown by the non-significant tendency of six-channel noise-vocoded sounds to be better recognised than one-channel noise-vocoded sounds.

In Experiment 6, different constellations of acoustic cues predicted participants' ratings on the different emotional scales. The results from this study suggest that a complex interaction of acoustic features is likely used in the perception of non-verbal emotional vocalisations. Fear and anger ratings were based on spectral and pitch information, surprise ratings on pitch and envelope cues, disgust from all three classes of acoustic information, and sadness and all of the positive rating scales on the basis of spectral and envelope cues.

The importance of different classes of acoustic cues clearly varies across different emotions: Although identification of many of the sounds relies heavily on pitch cues, recognition of amusement and anger sounds may rely more on other types of cues. In Experiment 5, amusement sounds were well recognised across all acoustic manipulations, indicating an important role for the amplitude envelope in the recognition of this emotion. This fits with the results from Experiment 6, where amusement ratings were predicted mainly on the basis of envelope cues. In Experiment 5, anger was better recognised from the six-channel noise-vocoded than the spectrally rotated stimuli, suggesting more of a role for gross spectral properties in the recognition of this emotion. Consistent with this, anger ratings were predicted mainly from spectral cues in Experiment 6.

Arousal and valence

In Experiment 6, multiple regressions with the acoustic cues of the sounds were used to predict participants' ratings on the arousal and valence scales in Experiment 2. Bachorowski (1999) has argued that vocal communication in speech sounds primarily signals the sender's arousal state, and only to a small degree their valence state. In line with her suggestion, the acoustic cues in the model in Experiment 6 explained 55% of the variance of the arousal ratings, but only 8% of the valence ratings. This pattern suggests that the perceived arousal, but not the valence, of emotional vocalisations can be well mapped out in terms of their acoustic features. However, the results from the PCA analysis of the ratings data in Experiment 2 suggest that valence is the dominant perceptual component — at least for these non-verbal emotional vocalisations: The principal component accounting for most of the variance (53.3%) correlated highly with the participants' valence ratings. The principal component correlated with arousal accounted for only 15.7% of the variance in the participants' ratings. It is of course possible that the acoustic features of the sounds that underlay the participants' valence ratings in Experiment 2 were not captured by this acoustic analysis, although a similar pattern has also been found in other studies (Laukka, 2004). One speculative explanation could be that valence represents a conceptual division rather than a perceptual one. It underlies participants' ratings of emotional stimuli without itself corresponding to any consistent set of acoustic cues. Further work will be needed to establish whether there is a set of acoustic cues that capture the valence aspect of emotional sounds.

In sum, the results from Experiments 5 and 6 show that acoustic cues play an important role in the perception of non-verbal emotional vocalisations. Experiment 5 demonstrated that removing acoustic cues impairs classification of emotional sounds. Pitch, pitch variation, and fine spectral detail seem to play important roles, although some emotions such as anger and amusement seem to rely primarily on other cues. Experiment 6 showed that acoustic cues can be used to classify emotional vocalisations using statistical models such as discriminant analysis. Participants' ratings of the sounds can predicted from the acoustic cues and a set of acoustic cues map onto perceived arousal, but valence does not correspond strongly to a set of acoustic cues.

Previous studies have identified pitch and aspects of pitch variation as important cues for emotion in speech (Murray & Arnott, 1993; Bänziger & Scherer, 2005). The findings from the current studies confirm a crucial role for pitch in the perception of emotion also in non-verbal vocalisations, although the set of cues that are used differs between the different emotions.

4. HOW DO THE HIMBA LAUGH? — AN INVESTIGATION OF VOCAL EXPRESSIONS OF EMOTION IN A PRE-LITERATE CULTURE

This chapter examines the issue of universality: do people communicate emotions using the same vocal signals regardless of their culture? Two studies test recognition in a sample of pre-literate adults of the Namibian Himba tribe, using stimuli produced by Western posers. These studies demonstrate that non-verbal vocalisations of emotions can be recognised cross-culturally. Specifically, Experiment 7 showed that Himba participants were able to reliably match emotional vocalisations to brief emotion scenarios. Female participants performed better than male participants. Experiment 8 replicated this finding using a task of same-different judgements. There was no main effect of participant gender in this task, although female participants performed significantly better than male participants in the recognition of anger sounds. Experiment 9 demonstrated that Western participants can recognise emotional vocalisations produced by Himba posers. In line with Elfenbein & Ambady's (2003) dialect account, performance was better when the poser and listener were from the same culture. The results of these studies provide evidence that the communication of emotions via non-verbal vocalisations is universal.

The universality of facial expressions of emotion

Central to the argument of basic emotions is the issue of universality: do people communicate emotions using the same signals regardless of their culture? Psychological universals are "core mental attributes that are shared at some conceptual level by all or nearly all non-brain damaged adult human beings across cultures" (Norenzayan & Heine, 2005, p. 763). Universals are of value to psychological research as they allow generalisations across populations with differing languages, cultures and ecologies. There is some disagreement on the best way of identifying psychological universals (Norenzayan & Heine, 2005; Russell, 1994), but most theorists assign particular importance to agreement between participants from different language groups and cultures that matches a stimulus to a specific concept (Ekman, 1994). In emotion research, universality has mainly been studied in the context of facial expressions of emotions. In other domains, research has addressed the universality of a number of psychological concepts including colour categories (e.g., Davidoff, Davies, & Roberson, 1999; Roberson, Davidoff, Davies & Shapiro, 2005), folk biology (Medin & Atran, 1999), children's Theory of Mind (Avis & Harris, 1991), and stereotypes about national characteristics (Terracciano et al., 2005).

Research in the late 1960s investigated whether people could recognise facial expressions of emotions regardless of culture. This work demonstrated that people from a visually isolated pre-literate culture, the Fore in New Guinea, were able to reliably match facial expressions of Western posers with emotion scenarios (Ekman et al., 1969). This finding showed that the facial cues utilised by Westerners to communicate emotional states were meaningful to members of a culture that had not been exposed to them before. This is important because it ruled out the possibility that signals of emotional communication are culture-specific, which was the dominant scientific view held at that time (e.g., Birdwhistell, 1970). These data showed that facial expressions of some emotions are shared across cultures, indicating that the basis for this system is universal. Also, facial expressions displayed by Fore individuals were reliably identified in terms of their emotional content by Western participants (Ekman and Friesen, 1971), thereby confirming this pattern. Ekman and colleagues' experiments provided crucial evidence of universality by demonstrating that facial expressions can be communicated to non-Western participants using Western posers and vice-versa.

More recently, a large number of studies have replicated and extended Ekman and colleagues' work. In 2002, a meta-analysis of 97 cross-cultural studies of emotional expression decoding (emotion recognition) confirmed their original conclusion that people of different cultures can identify facial expressions of the basic emotions at a level that is reliably above chance (Elfenbein and Ambady, 2002b). The meta-analysis found that recognition rates did not vary for specific emotions. This contrasts with Ekman and colleagues' early studies, where members of both Fore and Western cultures were unable to reliably distinguish surprise expressions from fear expressions for stimuli depicting members of the other culture. Ekman & Friesen (1971) speculated that this may be because, in this culture, situations eliciting fear would also be surprising. However, subsequent cross-cultural work has not tended to investigate pre-literate cultures, but rather compared emotion recognition across more similar cultural groups. Across this larger sample comparing across many cultures participants had no problems recognizing any specific emotion, and were better than chance for all of the original basic emotions (Elfenbein and Ambady, 2002b).

The dialect account of emotional communication

Investigations into the universality of emotional communication have tended to focus on the ability of participants to identify emotional expressions at a level that exceeds chance. However, in their meta-analysis Elfenbein and Ambady (2002b) found that in cross-cultural comparisons of recognition the group with the highest recognition scores tend to the cases where the participant and the poser were from the same culture. A number of different explanations have been put forward to explain this pattern of results. Matsumoto (1989) has suggested that in some cultures, such as the Japanese, understanding the emotion of another may be inhibited if that emotion is deemed to be socially disruptive. For example, Japanese participants may be less likely to identify a stimulus as angry, if anger is generally considered a socially inappropriate emotion in Japanese culture. Although this may be a useful account to explain differences in emotion recognition between North American and Japanese participants, it is unlikely that it could explain differences across the world.

Matsumoto has also pointed out that languages differ in their emotion vocabulary, that is, how many words they have to describe emotional states. He argues that the relative richness or poverty of a language to describe emotion could affect the processing of emotions, in speakers of those languages. For example, speakers of a language with a rich emotion vocabulary would be likely to process emotions more efficiently than others (Matsumoto & Assar, 1992). However, differences in recognition have been found between cultures that share the same language, such as England, North America, and Scotland (see Elfenbein and Ambady, 2002b).

In order to explain the in-group bias, Elfenbein and Ambady (2003) proposed a dialect account of emotional communication. This model builds on earlier accounts using universal emotional expressions based on general affect program, and display rules, conscious manipulations used to mask, accentuate or override expressions (e.g., Ekman, 1972). According to the dialect account, each cultural group also has a specific affect program, whereby members of that cultural group incorporate some minor adjustments to the original expressions. These, in combination with the display rules yield subtle differences in expressions between cultures (see Figure 4.1). The specific affect programs are acquired by social learning and so more exposure to a specific affect program would result in better recognition of emotional expressions from members of that culture.

According to Elfenbein and Ambady's (2003) dialect account, the processes involved in expression and perception of emotion are closely linked. Cultural differences can arise in two stages,

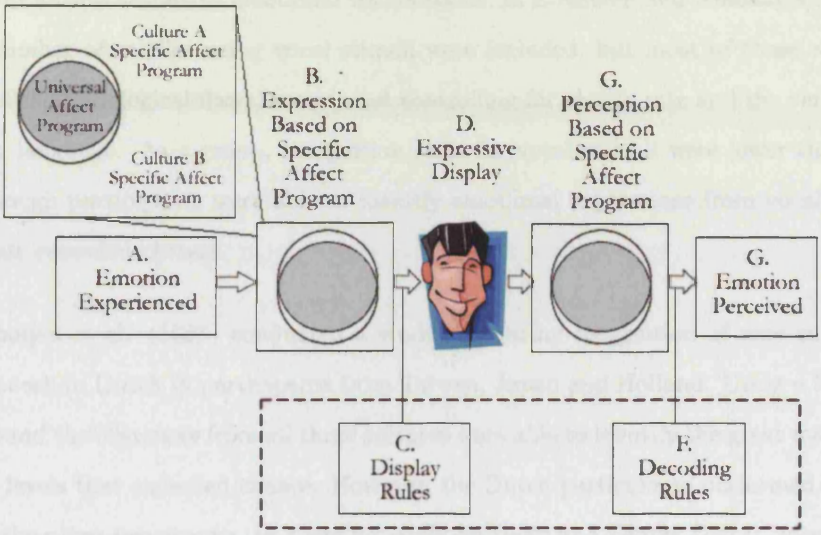
Culture A Specific Affect Program
Universal Affect Program
Culture B Specific Affect Program
B. Expression Based on Specific Affect Program
D. Expressive Display
G. Perception Based on Specific Affect Program
A. Emotion Experienced
G. Emotion Perceived
C. Display Rules
F. Decoding Rules

Fig. 4.1: Elfenbein and Ambady's (2003) dialect account of emotional communication. The universal affect programs are represented as grey, filled circles, with the culturally specific affect programs represented as the partially overlapping circles around them. Adapted from Elfenbein & Ambaday, 2003

as the information is filtered through the specific affect program both when expressing and perceiving an emotion. Although this model was developed mainly on the basis of data from facial expressions, it could also apply to vocal expressions of emotion. Just as smiles or frowns could be modified by affect programs, vocal expressions of emotions could be filtered through the same mechanisms, making emotional vocalisations from a member of the same culture easier to decode than vocalisations produced by an individual from a different culture.

Cross-cultural studies of emotional vocalisations

Not much work has been done to investigate whether vocal expressions of emotions are recognised cross-culturally. In an early study, Albas, McCluskey, and Albas (1976) investigated emotional speech in two groups of Canadians, Anglo-Canadians and Cree Indian Canadians. They used vocalisations in English and Cree that had been stripped of verbal contents by electronic masking. The researchers found that participants were better at identifying emotions communicated by a member of their own ethnic group. However, this study only included male participants and the stimulus set used was phoneme-based and thus similar to spoken language. This study is typical of early cross-cultural studies of emotional vocalisations; In Elfenbein and Ambady's (2002) meta-analysis a number of studies using vocal stimuli were included, but most of those were old and suffered from methodological flaws, such as not controlling for speech rate and the verbal contents used in each language. As a group, recognition rates for vocal stimuli were lower than for facial stimuli, although participants were able to identify emotional expressions from vocal expressions at a level that exceeded chance.

van Bezooijen et al. (1983) conducted a study comparing recognition of nine emotions from emotional speech in Dutch in participants from Taiwan, Japan and Holland. Using a forced-choice task, they found that listeners from all three cultures were able to identify the great majority of the emotions at levels that exceeded chance. However, the Dutch participants performed significantly better than the other two groups. In a similar study by Beier and Zautra (1972), speech segments in English were played to participants from the United States, Poland, and Japan. The segments expressed six different emotions and were of four different lengths, ranging from a single word to a sentence. Beier and Zautra found that as the length of the stimuli increased, the performance of the non-English speaking participants improved. Overall, the American participants performed better than the participants in the other groups, but this difference was negligible when the longest stimuli were used.

The studies by van Bezooijen et al. (1983) and Beier and Zautra (1972) are typical of studies investigating the cross-cultural communication of emotions, in that they examine the recognition of Western stimuli in participants in a number of different cultures. A recent study by Thompson and Balkwill (2006) used a different approach, where stimuli were generated by speakers of a number of different languages, and recognition stüdied in English speaking participants. Thompson and Balkwill collected semantically neutral sentences from speakers of English, German, Chinese, Japanese, and Tagalog (spoken in the Philippines), inflected with joy, anger, sadness and fear. They found that English-speaking participants were able to reliably identify all of the emotions from all languages, although the participants were most successful when presented with English stimuli, and least successful with Japanese and Chinese stimuli. The authors interpret their data as supporting Elfenbein and Ambady's (2003) dialect account.

In terms of cross-cultural studies of non-verbal expressions of emotion, a study by Scherer, Banse, and Wallbott (2001) tested recognition of non-verbal vocal expressions of anger, fear, sadness, joy, and neutrality. The study consisted of students from universities in seven European countries, the USA and Indonesia. The authors had originally intended to also include disgust, but pilot testing revealed that both recognition scores and ratings (how disgusted does this sound?) were low and so the disgust stimuli were excluded (see Scherer, Banse, Wallbott & Goldbeck, 1991 for similar problems with recognition of disgust sounds). They used sequences of six European (German, English, French, Italian, Spanish, and Danish) language nonsense-syllables that were free from semantic information, although they were made up from language-specific phonemes. They found an overall recognition rate of 66% and strong inter-cultural consistency of confusion patterns. This study reported that vocalisations expressing joy were the ones that participants had most problems categorising, with only 42% correctly recognised, compared to 66–76% for the other emotions. This may be due to the fact that they used only one type of positive emotional vocalisations. As noted previously, the stimuli used consisted of syllables from several Western-European languages, and were thus not entirely devoid of phonetic information; in fact the authors refer to their stimuli as "language-free speech samples" (p. 76). This may have biased the results in favour of the participants who spoke languages related to the Western-European languages used, as phonemes such as vowels vary depending on language family. This was indeed the case: the Indonesian participants, the only group whose native language was not related to any of the languages used to produce the stimuli, performed worst on the recognition task. Nevertheless, this study demonstrated that emotional prosody can be recognised across several cultures.

In order to avoid any biases from speech-like stimuli, the current studies used a set of stimuli that minimized any phonetic or syllabic content. This allows for an investigation more similar to the work carried out in the domain of facial expressions, in that the stimuli are truly non-verbal.

The aims of the current studies

These studies investigate whether non-verbal expressions of emotions could be recognised cross-culturally, using the two-culture approach as described by Norenzayan and Heine (2005). This method compares participants from two populations that are maximally different in terms of language, literacy, philosophical traditions etc. The claim of universality is strengthened to the extent that the same phenomenon is found in both groups. These studies also investigate whether Western participants listening to segments of vocalisations produced by non-Western posers are able to recognise the emotions expressed. Previous work using vocalisations has looked mainly at the recognition of Western stimuli in different cultural groups, and hence has not addressed the bi-directionality of the communication.

Experiment 7 — Matching non-verbal emotional vocalisations to emotion stories

In the first demonstration of the universality of emotional expressions, Ekman et al. (1969) used a task adapted from Dashiell (1927), originally developed for testing young children. The task involved telling the participant an emotional story whilst they were shown three photos depicting facial expressions of three different emotions. The participant was asked to select the picture in which the person's face showed the emotion portrayed in the story. The task is suitable for use with participants from a pre-literate culture as it does not rely on the translation of precise emotion terms and it can be used with pre-literate participants. This task was adapted for use with auditory stimuli for this study, with the participants being told an emotion story and subsequently choosing between two emotional sounds. In adapting the task from pictures to sounds, the number of response alternatives was reduced from three to two. This was done to avoid overloading the participants' working memory: Pictorial alternatives can be presented simultaneously, meaning that participants do not need to remember the response options. In contrast, sounds necessarily exist over time and hence cannot be presented simultaneously. The participant is required to remember the other response alternative(s) whilst listening to the current response option. There was a concern that with a population naïve to experimental testing, this kind of study design may challenge the participants' working memory. Consequently, two response alternatives were used instead of the three alternatives used in the study by Ekman et al. This study also aimed to examine gender effects in this sample, as previous research has found that females tend to perform better than men in the recognition of facial emotional stimuli in a pre-literate culture (Ekman & Friesen, 1971).

Methods

Stimuli

The stimuli were the same set used in Study 4 (9-way), thus selected on the basis of highest recognition scores in the pilot studies. Ten stimuli expressed each of the emotions anger, amusement, disgust, fear, pleasure, relief, sadness, surprise and achievement/triumph. The emotion contentment was excluded as these stimuli had been found to be systematically confused with sensual pleasure in previous experiments (see Experiments 1 and 2).

Participants

Twelve adult men and eleven adult women were tested. As the Himba do not count age, no age data are reported here. The subjects were each paid for their participation with one kilo of flour or sugar.

Design & Procedure

Each subject was tested individually in the presence of the experimenter and translator. The subject was told a short emotion story (see Appendix G) and allowed to ask questions to ensure that they had fully understood the story. The emotion stories were developed together with a local person familiar with the culture of the Himba people, who also acted as a translator during testing. When it was ensured that the subject had understood the story, two stimuli were played into the subject's headphones from an Apple iPod. The experimenter, but not the translator, wore headphones playing the same content as the subject's, to ensure the stimuli could be heard at a comfortable sound level. The participants' task was to choose which of the two stimuli matched the emotion in the story they had just heard and to report this to the translator who then relayed the answer to the experimenter. The subject was allowed to hear the sounds as many times as they required. This procedure was repeated for a male and female scenario for all of the nine emotions, with the genders and emotions presented in a random order for each participant. The target stimulus was of the same emotion as the story, and the distractor was of one of the other emotions, and any given trial contained stimuli produced by speakers of the same gender. The distractor stimulus was truly random and not systematically varied, save that it was never of the same emotion as the target stimulus. The order of the target stimulus and the distractor stimulus was random.

Results

Chi-square analyses of the data was carried out by grouping the responses into correct and incorrect responses, as there was no systematic variation of the distractor stimulus, and there was thus variability in the number of each category pairings. Overall, participants were significantly better than chance at matching the sounds with stories ($\chi_{(1)} = 65.0$, $p < 0.001$). When examining each emotion separately, chi-square analyses (Bonferroni-corrected for multiple comparisons) revealed

Tab. 4.1: Correct responses (%) of Himba participants matching emotional vocalisations using stories. Difference scores calculated as female greater than male participants' performance. Note: Chance level = 50%.

Emotion	Total	Males	Females	Difference
Achievement/Triumph	69.6	45.8	95.5	49.7 [2]
Amusement	91.3	87.5	95.5	8.0
Anger	58.7	54.2	63.6	9.4
Disgust	67.4	54.2	81.6	27.4 [1]
Fear	87.0	83.3	90.9	7.6
Pleasure	54.4	33.3	77.3	44.0[3]
Relief	84.8	83.3	86.4	3.1
Sadness	60.9	62.5	59.1	-3.4
Surprise	54.4	58.3	50.0	-8.3
Total Average	69.8	62.5	77.8	15.3

[1] Indicates $p < .05$

[2] Indicates $p < .01$

[3] Indicates $p < .001$

that participants' performance was significantly better than chance for sounds of amusement $\chi_{(1)} = 31.4$, $p < 0.001$), fear $\chi_{(1)} = 25.1$, $p < 0.001$), and relief $\chi_{(1)} = 22.3$, $p < 0.001$). The difference between participants' performance and chance for stimuli of achievement/triumph and disgust did not survive correction, although performance was quite high (achievement/triumph $\chi_{(1)} = 7.0$, $p < 0.01$ uncorrected, and disgust $\chi_{(1)} = 5.6$, $p < 0.025$ uncorrected). Participants were not significantly better than chance at matching sounds with stories for sadness, anger, surprise, or sensual pleasure. However, when data from only the female participants were considered, the performance for sensual pleasure stimuli was significantly better than chance ($\chi_{(1)} = 6.6$, $p < 0.025$).

A mixed ANOVA was performed to investigate the effect of stimulus category and speaker- and participant gender, on the recognition performance. Two within-subject factors were included, emotion and speaker gender, and participant gender was included as a between-subject factor. A main effect of emotion ($F_{(8,168)} = 5.4$, $p < 0.0001$) was found (see Table 4.1 for recognition rates of the different emotions).

A main effect of gender of the participants ($F_{(1,168)} = 6.3$, $p < 0.05$) was found (see Table 4.1), with females performing better on the task than males (on average 77.8% and 62.5% correct, respectively). In addition, there was a significant interaction between emotion and the gender of the participant ($F_{(8,168)} = 2.7$, $p < 0.01$), resulting from the differences in accuracy between the genders being more pronounced for some emotions than others. There was no main effect of the gender of the speaker, nor any interactions with that factor.

A one-way ANOVA was carried out to investigate gender differences for each emotion. There were significant differences in performance for achievement/triumph ($F_{(1,22)} = 20.4$, $p < 0.001$), disgust ($F_{(1,22)} = 4.9$, $p < 0.05$), and sensual pleasure ($F_{(1,22)} = 9.9$, $p < 0.01$); see Table 4.1.

Discussion

The results from this study demonstrate that non-verbal vocal expressions of emotions can be recognised cross-culturally. This lends support to Ekman's (1992b) suggestion that there is a set of positive basic emotions that are expressed vocally. Vocal expressions of amusement, fear and relief were reliably identified by the participants. Recognition of vocalisations of achievement/triumph and disgust performance was high, although recognition scores failed to reach statistical significance after multiple-comparison corrections. However, it seems unlikely that this pattern of results represents a true failure of the participants to associate the vocalisations to the emotion stories, considering the quality of the participants' performance (69.6% correct for achievement/triumph and 67.4% correct for disgust). Rather it seems this result reflects the lack of power of this data set: Each participant only yielded two data points per emotion, choosing between one correct and one incorrect response alternative for each female and male story scenario.

It is worth noting that the overall accuracy in this sample was lower than for Western participants. British participants recognised vocalisations at a level that exceeded chance for all emotions, in a forced choice task with ten options (see Experiment 2). The Himba participants had only two options in each trial and yet results failed to reach statistical significance for a number of emotions, despite positive trends. Although this is likely to in part be due to the low power of this analysis, this pattern was also found in Ekman and colleagues' original study. The researchers put this down to language barriers (i.e., testing via a translator) and task unfamiliarity (Ekman et al., 1969), factors which are likely to have contributed also in the current study.

Gender differences

In this study, female participants performed significantly better than male participants. Accuracy between male and female participants differed by up to almost 50% (for achievement/triumph sounds), although this was not the case for all emotions. Sounds of sensual pleasure were only reliably recognised by female subjects. The reason for this is unclear, though it may suggest that there is some difference in the way that Himba men and women express these emotions, with the female vocalisations being more similar to the Western vocalisations. Of course, both the male and female Himba would have heard vocalisations by both genders of their own culture, but it could be that it is easier to recognise a signal that is similar to one's own.

Other possible contributing effects to the higher performance of the female participants are higher motivation, better ability to concentrate or an actual underlying difference in emotion recognition ability. The lifestyle of the Himba may be a factor: the Himba women spend most of their time in the village with the other women, whereas the Himba men tend to take care of the cattle, which is more solitary. This could mean that the Himba women are more attuned to others' vocalisations as they have had more exposure to them, whereas the Himba men spend substantially less time with other people and hence have had less opportunities for this kind of interaction.

Gender differences in emotional expression recognition in non-Western cultures have been reported previously for facial expressions (Ekman & Friesen, 1971; Elfenbein, & Ambady, 2003; Hall & Matsumoto, 2004). Within-culture advantages for females in emotion recognition has been found both in studies using facial (Hall, 1978) as well as vocal speech stimuli (Toivanen, Vyrynen & Seppnen, 2005). Compared to male babies, female babies have been found to have a stronger innate preference for social stimuli, possibly giving the female infants an advantage in developing their superior perception of others' emotional signals (Connellan, Baron-Cohen, Wheelwright, Batki, & Ahluwalia, 2001).

Specific emotions: Anger, Sadness and Surprise

Sounds of anger, sadness and surprise were not reliably recognised by the Himba participants. This was true also for expressions for achievement and disgust stimuli, when the analysis was corrected for multiple comparisons. In the case of achievement/triumph, male participants performed at chance level whereas female participants were at ceiling, as mentioned. This has been discussed in the section on gender differences above. A likely reason that sounds of disgust were not reliably recognised was the low power in the current study. Each participant only gave two responses for each emotion, one for a male and one for a female scenario. The next study yielded more data points per participant and hence will address this issue empirically.

The Himba participants did not reliably recognise sounds of surprise in the current study. Previous cross-cultural work in preliterate cultures has also found that expressions of surprise are not reliably recognised (Ekman and Friesen, 1971), and some doubt has been expressed regarding whether surprise is a basic emotion (Reisenzein, 2000). The finding from this study adds support to the previous work showing that expressions of surprise are not recognised by members of preliterate cultures.

Regarding anger, one plausible explanation is that nonverbal vocalisations may not be typically used to express anger. It is possible that anger is preferably vocalised in speech sequences. Relative to other emotions, a person experiencing anger may be particularly keen to communicate the specific reason for their feelings, as anger tends to be caused by other individuals' actions (Scherer, 1997). The emotion story used for anger in Ekman's original study with the Fore was "he/she is angry, and about to fight" (Ekman & Friesen, 1971, p. 126), clearly something that would involve another person.

Using stimuli made up from European language nonsense-syllables, Scherer, Banse, and Wallbott (2001) found that the highest recognition rate in their study was for anger expressions. This discrepancy between the current study and Scherer et al.'s study may indicate that the recognition of vocal expressions of anger may be dependent on phonemic information.

The Himba participants were not able to recognise vocal expressions of sadness, even though it was clear that they did recognise facial expressions of sadness (personal observation). This could mean one of several things: either sadness is a universal emotion whose expression is exclusively conveyed via the face, or the stimuli used in the current study were inadequate. This cannot be established with the data from the current study and will be discussed further in the General Discussion of this chapter. In summary the current study provides some support for the existence of a set of cross-culturally recognisable vocal signals for achievement/triumph, amusement, disgust, fear and relief.

Experiment 8 — Differentiating classes of non-verbal vocalisations of emotions

The task used in Study 7 relied to a high degree on the suitability of the emotion stories used. The participants' poor performance in recognising vocalisations of certain emotions could have been due to inadequate emotion stories rather than having been caused by cross-cultural differences in emotional vocalisations. In order to provide complementary data that did not involve emotion stories, a second task was designed, modelled on a test used with another non-literate group, children with dyslexia (Richardson, Thomson, Scott & Goswami, 2004). In Richardson et al.'s study, the participant would hear two sounds with either the same or different fundamental frequency, and their task was to respond whether they thought the sounds were the same or different. However, in the current study the point of interest was not simple acoustic discrimination, so the task was adapted so that each trial would contain two different sounds, with half the trials containing sounds expressing the same emotion and the other half of the trials containing sounds expressing different emotions. The aim of the study was to establish whether the findings from Study 7 were replicable using a different paradigm, and whether the emotions that were not reliably identified in Study 7 would be reliably recognised in a study with increased power.

Methods

Stimuli

The same stimuli as in Study 7 were used.

Participants

Sixteen participants (8 male, 8 female) were included in the study. The participants had not participated in Study 7.

Design & Procedure

Each subject was tested individually in the presence of the experimenter and translator. Two stimuli were played into the subject's headphones from an Apple iPod. The participants' task was to judge whether the two stimuli expressed the same emotion or two different emotions. As in Study 7, the participant reported their judgement to the translator who then relayed the answer to the experimenter, and the participant was allowed to hear the sounds as many times as they required. The experimenter, but not the translator, wore headphones playing the same content as the participant's, to ensure that the stimuli could be heard. In half of the trials, the two stimuli expressed two random different emotions, and in the other half of the trials the stimuli expressed the same emotion. The two stimuli were always produced by different speakers of the same sex.

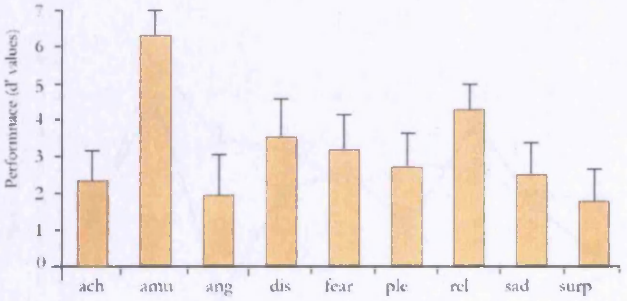

Fig. 4.2: Performance of Himba participants on Same-Different task (d' scores). Note: Zero is chance performance; higher d' value denotes better performance.

Results

Chi-square tests using the raw data revealed that participants performed significantly better than chance overall ($\chi_{(8)} = 155.5$, $p < 0.001$, and also for all of the emotions individually, except surprise. The chi-values for each emotion are displayed in Appendix H.

Investigation of the data revealed that participants showed a response bias, in that they tended to respond "different" more often than "same". Consequently, the participants' performance was analysed using signal detection analysis. The equation from Macmillan and Credman (1996) was used (see below). The performance of the participants for each emotion using d' scores is shown in Figure 4.2.

$$C = 0.5[2(H) + z(F)]$$

$$d' = z(H) - z(F)$$

C = response bias

d' = correct score (adjusted for response bias)

H = hits

z = normal distribution function

F = false alarms

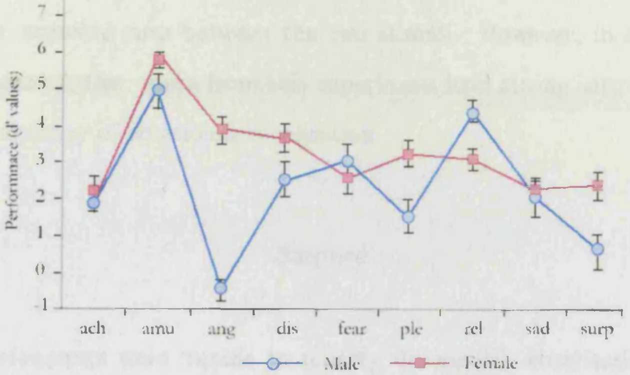

Fig. 4.3: Performance (d') for male and female Himba participants in the same-different judgment task for each emotion.

A repeated measures ANOVA of the d' scores with emotion as a within-subjects measure and gender as a between-subjects measure, revealed a significant effect of emotion ($F_{(8,120)} = 2.5$, p < 0.05), indicating that the participants' performance differed between emotions. There was no significant effect of gender and no significant interaction between gender and emotion. However, a visual inspection of the data suggested possible differences in performance for individual emotions between the gender groups, notably for anger (see Figure 4.3).

A one-way ANOVA was carried out, testing for differences between male and female participants for each emotion. There was a significant difference in performance for judgements of anger stimuli ($F(1, 14) = 8.8$, p < 0.01), but no significant difference for any other emotion. In the case of anger, females performed significantly better than men (d' means 3.9 and -0.4, respectively).

Discussion

The results from this study indicate that the Himba participants were able to make accurate judgements about whether two emotional vocalisations expressed the same emotion or two different emotions. This was true overall, and also for all emotions except surprise. This pattern of results indicates that this task was easier than the task used in Experiment 7, as participants performed reliably above chance for a larger set of emotions. However, this may be because the participants could have based their judgements on the degree of overlap in the acoustic properties of the two sounds; Two vocalisations expressing the same emotion are likely to be more acoustically similar than two vocalisations expressing two different emotions. In order to minimize the use of this strategy, the two stimuli heard within a single trial were never identical, and were always produced by two different speakers. Nevertheless, it remains possible that participants relied to some extent on overlap of simple acoustic cues between the two stimuli. However, in conjunction with the results from Experiment 7, the data from this experiment lend strong support to the hypothesis of cross-cultural consistency of emotional vocalisation.

Surprise

Importantly, the participants were unable to reliably distinguish vocalisations of surprise from other emotions. This is consistent with the results from Experiment 7, where participant did not reliably pair vocalisations of surprise with the surprise story. This suggests that the inferior performance with surprise sounds in Experiment 7 was unlikely to have been due to the use of an inappropriate emotion story. Rather, the data from Experiments 7 and 8 taken together indicates that the Himba participants do not recognise emotional vocalisations of surprise produced by Westerners. Previous cross-cultural work has found that members of non-literate cultures are unable to distinguish facial expressions of surprise from expressions of fear (Ekman & Friesen, 1971), indicating that surprise may be not a universally recognizable emotion when communicated using the face or voice. It is possible that members of the Himba or other non-literate cultures would reliably identify expressions of surprise if communicated by full-body posture or speech segments; this may be a question for future cross-cultural studies to explore. Also, it would be important to establish whether Himba participants could recognise vocalisations of surprise posed by members of their own culture.

Gender effects

Although there was no overall effect of gender on performance, female participants performed significantly better than male participants for anger stimuli. In Experiment 7 anger sounds were one of the most poorly recognised stimulus types. However, in that study there was no gender effect on performance, with both male and female participants performing at chance levels. This may have been due to the relative difficulty of the two tasks; possibly this might suggest that the anger scenarios used in Study 7 were inappropriate. Another alternative is that this finding reflects the lack of power in Study 7. It is worth pointing out that the participants in Experiment 8 could have based their judgements on acoustic rather than emotional properties of the sounds. They were asked to focus on the emotion communicated, but the possibility that they may have relied primarily on acoustic cues can not be excluded. Why female, but not male participants would be able to distinguish anger sounds is not clear, but it may have been that the female participants were better able to utilize the acoustic cues.

In sum, this study demonstrates that Himba participants are able to accurately judge whether two different emotional vocalisations belong to the same emotion category, for all emotions except surprise.

Experiment 9 — Western recognition of Himba emotional vocalisations

Previous work investigating the cross-cultural reliability of emotional expressions has primarily tended to study the recognition of Western stimuli in non-Western participants (e.g., Ekman et al., 1969; Scherer, Banse, & Wallbott, 2001). Thus studies have generally neglected the bi-directionality of emotional communication, that is, the investigation of Western recognition of non-Western emotional stimuli (for an exception see Ekman & Friesen, 1971). However, to address the issue of whether emotional expressions are universal, it is necessary to examine recognition in both Western and non-Western participants, of stimuli produced by both Western and non-Western speakers (Matsumoto, 2002). One exception in the area of vocal communication of emotions is a recent study by Thompson & Balkwill (2006). They examined recognition of emotional speech in two Indo-European (English and German) and three Asian languages (Chinese, Japanese, and Tagalog). Using a forced-choice task with semantically neutral sentences inflected with joy, anger, sadness and fear, they showed that English-speaking participants were able to recognise all of the emotions in all of the languages.

This study examines the recognition by Western participants of emotional vocalisations produced by non-Western speakers. Using a more extensive set of emotions than Thompson and Balkwill (2006), it presents a more challenging task. As a forced-choice task was used, equivalent to Experiment 3, this allows for a comparison of the relative ease of recognition from stimuli produced by speakers from the participants' culture or another culture. Similarly to previous studies, non-verbal vocalisations of emotions were used. This avoids confounding differences in language with differences in culture, and these types of signals are also thought to be perceived as more reliable indicators of emotion (Johnstone & Scherer, 2000).

Methods

Stimuli

Vocal expressions of emotion were collected from 5 male and 6 female Himba participants. The speakers were recorded outdoors with a video camera, having been presented with appropriate scenarios for each emotion (see Appendix G). None of these participants had heard the Western stimuli, and no explicit guidance was given as to the sort of sounds the speakers should generate, that is, the speakers were not given exemplars to mimic, and they were instructed to avoid any 'verbal' items. Each speaker produced between one and four vocalisations per emotion. The sound from the resultant video clips was stripped off and the sound files saved as .WAV files. This resulted in 169 stimuli, with 17 expressing achievement/triumph, 21 expressing amusement, 16 expressing anger, 22 expressing disgust, 25 expressing fear, 13 expressing sensual pleasure, 18 expressing relief, 13 expressing sadness, and 24 expressing surprise.

Subjects

Twenty native English speakers (10 male, mean age 22.6 years) took part in the study. They were recruited from the UCL Psychology subject data base.

Design & Procedure

Design and procedure were identical to Experiment 3, except that each participant heard 169 stimuli rather than 90. This was a 9-way forced choice task in which the stimuli were played in a random order.

Results

The six best recognised stimuli of each emotion (from the experimental data) were included in the analysis. This was done in order to exclude stimuli of a low standard, both in terms of emotional expressivity and sound quality. Participants were highly successful at identifying the vocalisations (see Table 4.2).

Tab. 4.2: Recognition by Western participants of emotional vocalisations produced by Himba speakers (%), and chi-values for each emotion category. All chi-values Bonferroni corrected for multiple comparisons.

Emotion	Recognition	Chi
Achievement/Triumph	50.8	191.7^3
Amusement	93.3	821.4^3
Anger	40.8	107.3^3
Disgust	92.5	804.8^3
Fear	45.8	146.5^3
Pleasure	46.7	153.6^3
Relief	62.5	320.9^3
Sadness	80.8	590.6^3
Surprise	53.3	216.6^3

^{3} Indicates $p < .001$

Chi-square analyses of the data revealed that participants' performance was significantly above chance for all emotion categories (see Table 4.2). Notably, there is substantial variation in the participants' performance, with performance highest for amusement and disgust sounds and lowest for anger and fear sounds. A one-way ANOVA with gender as a between-subjects effect and emotion as a within-subjects effect revealed no effect of participants' gender for any emotion category.

In order to test Elfenbein & Ambady's (2002a; 2002b) prediction that recognition is superior for emotional stimuli produced by individuals from the same cultural group, a repeated measures ANOVA comparing Western participants' recognition of stimuli produced by Himba and Western speakers was carried out, using the recognition data from Study 3 as a comparison. The test used proportion of correct scores and included emotion as a within-subject factor with nine levels, and stimulus type as a between-subjects factor. The analysis yielded a significant main effect of stimulus type ($F_{(1,38)} = 56.1$, $p < 0.0001$), a main effect of emotion ($F_{(8,304)} = 18.2$, $p < 0.0001$), and a significant interaction ($F_{(8,304)} = 6.4$, $p < 0.0001$). Overall, the participants were more accurate when judging stimuli produced by the British as compared to the Himba speakers (see Figure 4.4). A one-way ANOVA tested this relationship for each emotion category, showing that performance differed for the two stimulus types for all emotions except amusement and relief (see Appendix I). For these two emotions, participants were marginally better at recognizing the Himba stimuli, although this difference was not significant. As previously, participants' performance varied by emotion, with some emotional vocalisations being easier to identify than others (see Figure 4.4). Also, the difference in performance between the two stimulus types varied by emotion (see Figure 4.4).

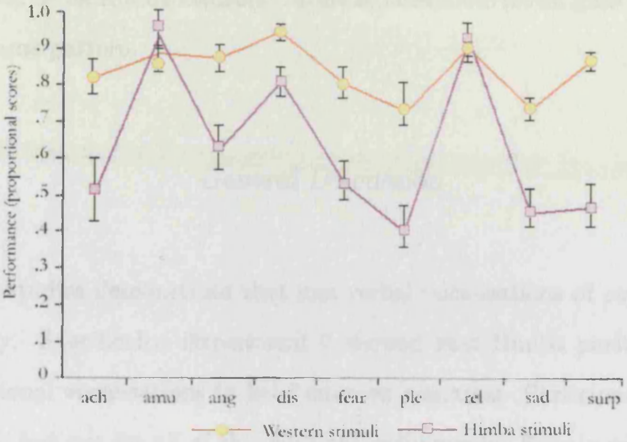

Fig. 4.4: Proportion correctly classified stimuli by emotion category for Himba and Western stimuli by British participants.

Discussion

The results from this study indicate that Western participants can recognise emotional vocalisations produced by non-Western participants, at a level reliably above chance for all emotion categories. There was substantial variation in the participants' performance, with performance ranging between 93% (amusement) to 40% (anger). This is the first demonstration that non-verbal emotional vocalisations can be communicated cross-culturally in both directions, between members of Western and non-Western cultures.

The dialect account

This data set together with the data set from Study 3 also allowed for a test of Elfenbein & Ambady's (2002a; 2002b) dialect account. This model predicts that recognition would be superior for emotional stimuli produced by individuals from the same cultural group. This hypothesis was confirmed: the Western participants were more accurate when judging stimuli produced by the British as compared to the Himba speakers. Although this was true overall and for most of the emotion categories individually, there was no difference in recognition performance for amusement and relief stimuli. This was likely because recognition for the Himba stimuli of these two emotions was so high: apart from disgust, these were the two best recognised emotions. This pattern likely reflects a ceiling effect for these emotion categories. In sum, this data set supports Elfenbein and Ambady's dialect account: The participants were better than chance at recognizing all stimulus types, but better for stimuli produced by members of their own culture. This study did not allow this model to be tested using Himba listeners: future studies could investigate whether non-Western listeners show the same pattern.

General Discussion

The results of these studies demonstrate that non-verbal vocalisations of emotions can be recognised cross-culturally. Specifically, Experiment 7 showed that Himba participants were able to reliably match emotional vocalisations to brief emotion scenarios. Participants' performance was above chance overall, but not for all of the emotions individually. Female participants performed better than male participants, especially for achievement/triumph and sensual pleasure. Experiment 8 replicated the basic finding from Experiment 7, using a same-different task. Participants were better than chance overall and for each of the emotions except surprise, in a task of same-different judgements. There was no main effect of participant gender, although female participants performed significantly better than male participants in the recognition of anger sounds. Experiment 9 demonstrated that Western participants can recognise emotional vocalisations produced by Himba posers. In line with Elfenbein & Ambady's dialect account (2002a; 2003), recognition of these stimuli was poorer than for stimuli produced by Western posers.

Gender effects

In Ekman & Friesen's original study of recognition of facial signals of emotions in a pre-literate sample, females performed better than males. This finding was replicated using vocal stimuli in Study 7. This pattern was especially strong for recognition of sounds of achievement/triumph and pleasure. This overall gender advantage was not found in Study 8, although female participants performed significantly better than male participants for anger stimuli. What is causing this pattern of results is not entirely clear. Study 9 showed that recognition in Western participants is not significantly affected by gender, suggesting this effect to be specific to non-literate cultures.

It could be suggested that the Himba women were in some way more Westernised than the men. However, this explanation does not seem satisfactory in light of the Himba lifestyle. To the extent that any Himba had come in contact with any Western culture, the men would be more likely than the women to have been exposed to Western influence. The gender roles are very strong in a traditional culture such as the Himba, and the women tend to spend their time in the village. The men look after the cattle and move across large areas of the desert, and a few of them would have visited the regional capital, Opuwo (population approximately 7000 people). These Himba men are very unlikely to have been exposed to any Western media or culture, although they may have seen a very small number of Western tourists. In any case, the degree of Western influence cannot explain the finding of females outperforming males, as this should have resulted in the opposite outcome.

A number of other factors could be suggested to account for the female participants' superior performance in Study 7, but in light of the lack of a main effect of gender in Study 8, caution must be exercised; also these explanations are necessarily speculative. One possibility is that the female participants were better at abstract thinking. In Study 8 the task could be carried out largely by matching stimuli on the basis of acoustic features (although this method was discouraged by the instructions emphasising the emotional content of the sounds). In Study 7, more abstract thinking was required on the part of the participant, as they had to imagine the emotional state of the person in the scenario. It may be that this task relied more heavily on some underlying ability — such as abstraction or Theory of Mind — that the female participants were more proficient in. There is of course also the possibility that this difference reflects an actual difference in ability. Himba women tend to spend more time with other members of their group and may therefore be more attuned to others' vocalisations. This does seem to contradict the lack of a gender effect in Study 8, but, as noted above, this finding could possibly be explained by the use of an alternative strategy in this study. Additional data is needed to establish whether this pattern of results reflects a true gender difference in emotion perception, or whether it is caused by differences in other underlying abilities such as abstract thinking.

Specific emotions

A number of emotions were consistently recognised at above-chance levels: Amusement, fear and relief were reliably identified by the participants in all three studies. Vocalisations of disgust were reliably recognised in Studies 8 and 9, but the results failed to reach statistical significance in Study 7. This is likely due to be due to the low power in Study 7, although this interpretation would need to be confirmed by additional data with a larger sample.

Sounds of achievement/triumph and sensual pleasure were reliably classified in Studies 8 and 9, but only female participants were able to match the sounds to the emotion scenarios in Study

7. A qualitative consideration of the Himba stimuli in Study 9 presents a possible explanation: there were large differences in the types of sounds produced by the male and female posers for the emotions achievement/triumph and sensual pleasure. The female sounds seem acoustically more similar to those of the Western stimuli (note that neither the male nor female participants producing the sounds had heard the Western stimuli) for both these emotions. No doubt all Himba participants would have heard sounds produced by both male and female Himba, but nevertheless it seems possible that individuals base their judgements to some extent on a comparison with the sounds that they themselves tend to make. This could potentially explain why women but not men could match sounds of achievement/triumph and sensual pleasure to emotion scenarios in Study 7. In Study 8 participants could have relied on judgements of acoustic cues to a larger degree.

Anger sounds were not recognised at a level that exceeded chance in Study 7, and only by female participants in Study 8. One plausible explanation is that non-verbal sounds may not be typically used to express anger, and that the female participants in Study 8 relied on acoustic cues to perform the task. Anger may be preferably vocalised in speech sequences. A previous cross-cultural study using language-free speech samples found a high recognition rate for anger expressions (Scherer, Banse, & Wallbott, 2001). This discrepancy between the current study and Scherer et al.'s study may indicate that the recognition of vocal expressions of anger may be dependent on phonemic information (see also Chapter 5).

Participants were unable to reliably distinguish vocalisations of surprise from other emotions, or to match surprise sounds to the appropriate scenario. This suggests that the Himba participants did not recognise emotional vocalisations of surprise produced by Westerners. Previous cross-cultural work has found that members of non-literate cultures are unable to distinguish facial expressions of surprise from expressions of fear (Ekman and Friesen, 1971), indicating that surprise may be not a universally recognizable emotion. This contrasts with the results from Study 9, where Western participants were able to reliably recognise vocalisations of surprise produced by Himba posers. Although surprise was one of the least well recognised emotions, the stimuli were recognised at a level that exceeded chance. In general, the Western participants performed better than the Himba participants, likely due to higher familiarity with psychological testing. It may be that the discrepancy in recognition of surprise stimuli between the two groups is a reflection of the generally poor performance of the Himba participants, and the relatively low power of Studies 7 and 8.

The dialect account

In Study 9, Western participants were able to recognise emotional vocalisations produced by Himba posers at a level that reliably exceeded chance. In line with the predictions generated from Elfenbein & Ambady's dialect account (2002a; 2002b; 2003), an in-group bias was found: recognition was superior for emotional stimuli produced by individuals from the same cultural group. The Western participants performed better when tested with stimuli produced by Western posers, than when tested with the Himba stimuli on the same task. It is worth noting that the participants in either case did not know the cultural group of posers of the sounds they were hearing. This pattern of in-group bias was true for the data as a whole and for most of the emotions individually. In the cases of amusement and relief sounds, performance did not differ between Himba and Western stimuli. This was likely due to a ceiling effect for these emotion categories. This study did not allow this model to be tested using Himba listeners: future studies could investigate whether non-Western listeners show the same pattern.

Universal mechanisms

These data constitute the first empirical demonstration that non-verbal vocalisations of emotion are universal. It is worth noting that these data do not speak to the issue of what causes this pan-cultural association between certain emotion concepts and their specific vocalisation patterns. In the context of facial expressions, Ekman and Friesen noted that "universals in facial behaviour with emotion can be explained from a number of nonexclusive viewpoints as being due to evolution, innate neural programs, or learning experiences common to human development regardless of culture" (p. 128, 1971), and this may also be true in the case of vocal signals of emotion. Teasing apart the different accounts of how these associations form will be a difficult task, likely involving studies of young infants and non-human animals (Norenzayan & Heine, 2005). Scherer, Banse and Wallbott have suggested the possibility of "universal coding relationships, based on emotion-specific physiological patterning affecting voice production" (2001, p. 89). This hypothesis would be difficult to test in a true cross-cultural setting, as the measuring of participants' physiological states in cultures such as the Himba would be near enough impossible. A comparison between participants from a number of cultures where access to electricity and controlled laboratory settings is available, may nevertheless provide a useful starting point.

Summary

These studies demonstrate that non-verbal vocalisations of emotions are communicative tokens that can be recognised by members of vastly different cultures. Members of the pre-literate Himba tribe were able to match emotional vocalisations to appropriate scenarios and to vocalisations of the same emotion. Western participants were able to recognise emotional vocalisations produced by Himba posers. In line with Elfenbein & Ambady's dialect account (2003), recognition was superior for emotional stimuli produced by individuals from the same cultural group. In sum, these data suggests that emotional vocalisations are universal.

5. THE COMMUNICATION OF EMOTIONS IN SPEECH

This chapter consists of three experiments using emotional speech stimuli. Experiment 10 addresses the issue of whether emotions can be recognised in speech using a forced-choice task. The data from this study show that naïve listeners are able to reliably identify emotions in speech, although performance is less accurate than with non-verbal emotional vocalisations. Experiment 11 employs a forced-choice task with acoustically manipulated sounds, to examine the contribution of different acoustic cues in the recognition of emotions in emotional speech stimuli. The results show a pattern broadly similar to that found with non-verbal stimuli in Experiment 5, although there is a somewhat stronger role for broad spectral cues in emotion recognition from speech stimuli. Experiment 12 uses stimuli at ten levels of noise-vocoding between one and 32 channels within three tasks: speech intelligibility, emotion recognition, and speaker differentiation. The results show that participants' performance improves with increasing numbers of channels in all three tasks, but at different rates. In sum, this chapter shows a number of similarities between the perception of emotional speech and non-verbal vocalisations but points towards differences in the perception of different types of information, such as verbal content and speaker identity, in emotional speech.

There has been an increased interest in the communication of emotion in the voice in recent years. With very few exceptions, this interest has focused on emotion conveyed in speech; more specifically, the para-linguistic aspects of speech, such as prosody. A number of studies have examined whether naïve listeners can infer the speaker's emotional state from emotional speech, mostly using forced-choice tasks or rating scales. In an early review, Scherer examined approximately 30 studies of emotions in speech carried out up until the early 1990s. He found that emotions were recognised at approximately 60% accuracy (Scherer, 1989). A more recent review yielded a similar result, with average recognition at 55–65% accuracy (Scherer, 2003). Scherer pointed out that accuracy varies greatly between emotional classes and that the overall recognition rate is somewhat lower than the average level found for facial expressions of emotions, which tends be around 75% (for a review of recognition of emotions from facial expressions see Ekman 1994). Nevertheless, it is now well established that naïve listeners are able to infer emotional states from emotionally inflected speech.

Emotional speech stimuli

As in all studies of emotional communication, the stimuli used are of great importance. A number of ways to produce emotional speech samples have been used: Some studies have used emotional speech taken from real conversations, some have used synthetically produced speech, and others have induced emotions in speakers and used the resultant emotional speech (see Laukka, 2004). However, most current studies use the "standard content paradigm" where the same verbal material is repeated with a number of different emotional infusions (Davitz, 1964). This has the advantage of keeping the verbal content constant across emotions and so avoids the listener inferring the speaker's emotional state from the speaker's choice of words. The main drawback of this methodology is that it relies on non-spontaneous speech, that is, posed emotional expressions. Considering the problems of reliably inducing speech in the laboratory and the lack of control of verbal contents in spontaneous speech samples from real settings, the standard contents paradigm is generally thought to be superior, and that is the methodology employed to produce the stimuli used in this chapter.

As in studies with facial expression stimuli, studies of emotional speech have tended to preselect stimuli on the basis of pilot data from recognition studies. Exceptions are studies that have matched stimuli for recognition (e.g., Scott et al., 1997), or used all stimuli produced in a set (Juslin & Laukka, 2001). The majority of studies have used emotional speech stimuli pre-selected for best recognition rates, yielding recognition rates of around 55–65% (see Scherer, 2003).

Scherer's component process theory

There is a dearth of theories on the communication of emotions in the voice. The one available theory is Scherer's component process model (Scherer, 1986; 2001; 2003). This theory is based around the idea that emotions affect our physiology, which in turn affects our vocal apparatus. According to Scherer's theory, emotions are the results of a series of appraisals or stimulus evaluation checks, which evaluate the emotional stimulus in terms of a number of aspects, such as novelty and pleasantness. Each stimulus evaluation check has a specific physiological effect which then in turn affects the vocal production system. Scherer gives the example of a stimulus being appraised as dangerous and requiring action. This would cause the somatic nervous system to increase muscle tension and the fundamental frequency of the voice would increase. At the same time salivation would decrease, contributing to the effect of high-pitched vocal output. Based on his model, Scherer has made an extensive set of predictions of acoustic properties expected in a number of emotions, such as sadness, anger and happiness. These predictions remain largely untested, largely due to the difficulty of measuring many of the acoustic parameters Scherer specifies (but see Grandjean & Scherer, 2006).

The aims of this chapter

This chapter is an investigation of communication of emotions in speech. It aims to examine whether the ten emotions studied in non-verbal vocalisations can be reliably identified in speech sounds. Experiments 10 and 11 also include the emotion "happiness", to examine the recognition of this broader emotion term as compared to specific positive emotions such as amusement and relief. In addition, Experiment 10 allows for a comparison of recognition of emotions in speech and non-speech stimuli, both in terms of overall accuracy and ease of recognition for individual emotions. Experiment 11 is a study of the role of acoustic cues in the recognition of emotions in speech. The acoustic manipulations employed in Experiment 5 are used here to allow for a comparison across stimulus types. Experiment 12 consists of three tasks with noise-vocoded speech stimuli. Listeners were required to identify the verbal contents of the sounds, the emotion, and to determine whether the stimuli were produced by the same or different speakers. This was done to investigate whether these judgments rely on the same acoustic information.

Experiment 10 — Can we communicate emotions via paralinguistic cues in speech?

In contrast to the lack of work investigating non-verbal vocal signals of emotion, there has been considerable interest in emotional communication in speech. Much of this work has attempted to establish whether naïve listeners can infer the emotional state of the sender from emotionally inflected speech. By now an extensive set of studies has shown that emotions can be reliably communicated in speech (for reviews see Juslin & Laukka, 2001; 2003; Scherer, 1986; 2003). For example, a recent review reported that listeners recognised emotions from speech with around 55–65% accuracy, much higher than would be expected by guessing (Scherer, 2003).

There are large differences in the ease of recognition for different emotions. According to a recent review by Scherer, sad or angry speech tends to be best recognised, followed by fearful speech (Scherer, 2003). This is consistent with another review of emotional speech by Juslin and Laukka (2003), who also found that anger and sadness expressions were best recognised. Recognition rates for disgusted and happy stimuli have tended to be remarkably low, and in at least one study disgust was not included as recognition in piloting stages was too poor (Scherer, Banse & Wallbott, 2001). Recognition scores for happiness have also tended to be much lower in studies using speech, whereas studies using face stimuli generally find ceiling effects for recognition of happiness (Scherer, 2003).

The aims of this experiment

This experiment was aimed at establishing whether naïve listeners could identify expressions of positive and negative emotions correctly from emotionally inflected speech. Although there is now a wealth of evidence suggesting that emotions can be reliably inferred from speech, in this study the production of the speech stimuli closely mirrored that of the non-verbal stimuli (Chapters 2–4) to enable a comparison of emotion recognition from verbal and non-verbal vocalisations of a wider range of emotions. In addition, stimuli expressing the broader emotion category "happiness" were included in order to examine this study in the context of previous work with emotional speech, as other studies have tended to use a single positive emotion category.

Method

Stimulus preparation and pilot

The verbal expressions of emotion were collected from the same two male and two female native British English speakers that had produced the non-verbal stimuli. The scenarios, emotional labels, and anechoic chamber used for recording were the same as those used for the previous recording session (see Appendix A). In addition to the ten categories from Experiment 2, the category "happiness" was added to facilitate comparison to previous work. The example given for happiness was "things go better than expected". Each speaker produced 10 speech stimuli per emotion category in the form of spoken three-digit numbers. The numbers were randomly generated by the experimenter and were the same across speakers and emotions[1]. The resultant 440 sounds were digitised at 32kHz.

All the stimuli were then pilot tested on nine participants, who performed a forced-choice task, procedurally identical to the main study (see below). A test set was chosen by selecting the ten best recognised stimuli of each emotion. The average accuracy rate was 79%, with emotions ranging from 41.1% for happiness to 100% for sadness. All speakers were represented in the stimuli of eight or more emotions, and all emotions but one (amusement) included stimuli from three or more speakers. Examples of the stimuli are available on a CD in the additional materials.

Participants

Twenty-two British English speaking participants (6 male, mean age 21.5 years) participated in this study. The participants were recruited from the University College London Psychology department participant database.

Design & Procedure

The design and procedure were identical to that of Experiment 2, except that the response options included the response alternative "happiness" in addition to the other emotional labels. The example given for happiness was "things go better than expected". The keys 0–9 were used for the response options, and the additional key "h" was used for the "happiness" label. In all other respects, the labels and sentences and instructions used were identical to those used in Experiment 2 (see Appendix A).

Results

Performance on the forced-choice task with speech stimuli

The listeners categorised the emotions with above chance accuracy (see Table 5.1). For each stimulus type except happiness, the most frequent response was the appropriate category. For happiness

[1] The numbers used were 847, 143, 707, 373, 545, 707, 847, 951, 137, and 545.

Tab. 5.1: Recognition of emotion in speech in an 11-way forced choice task. Correct categorisations in bold. Data are in percentages (%). Horizontal lines add to 100.Note: Ach = achievement/triumph, Amu = amusement, Ang = anger, Con = contentment, Dis = disgust, Hap = Happiness, Ple = sensual pleasure, Rel = relief, Sad = sadness, Surp = surprise.

Stim	Response										
	Ach	Amu	Ang	Con	Dis	Fear	Hap	Ple	Rel	Sad	Surp
Ach	**55.2**	3.9	1.7	1.7	0.4	0.4	26.5	2.6	3.9	0.0	2.6
Amu	0.0	**82.2**	0.0	0.4	0.0	0.0	9.1	2.6	0.9	0.4	4.3
Ang	0.4	0.9	**90.0**	1.7	6.5	0.0	0.4	0.0	0.0	0.0	0.0
Con	4.8	0.0	0.4	**69.1**	2.6	0.0	5.7	3.0	9.6	4.3	0.0
Dis	4.3	1.3	2.2	3.9	**56.5**	1.3	1.3	1.7	3.9	3.5	18.3
Fear	2.2	1.7	0.4	0.0	0.4	**79.1**	2.6	1.3	2.6	7.4	1.7
Hap	17.0	2.6	0.9	36.1	1.7	0.4	**27.4**	3.5	6.5	0.4	3.0
Ple	1.7	0.4	0.0	6.1	3.9	4.8	1.3	**54.8**	7.4	17.4	2.2
Rel	3.0	2.2	0.0	14.8	3.0	0.9	3.5	11.7	**55.7**	3.5	1.3
Sad	0.0	0.0	0.0	0.4	0.9	1.3	0.0	0.9	0.9	**94.3**	1.3
Surp	0.4	1.3	0.4	1.7	3.5	2.6	0.4	0.0	1.7	0.9	**87.0**

stimuli, the sounds were more often labeled contentment than happiness. Due to technical problems there was a small number of missing data points (10 responses missing). Chi-square analyses of the data revealed that participants were significantly above chance for all stimulus categories; chi-square values for each emotion category are given in Appendix J. Accuracy was highest for sadness (94.3%) and anger (90%). Happiness was by far the least well-recognised emotion, with only 27.4% of stimuli identified correctly.

Happiness sounds were confused with contentment (36.1% of responses) and achievement/triumph (17%). Other common confusions within the positive emotions were achievement/triumph sounds labeled happiness (26.5%), amusement sounds labeled happiness (9.1%), and relief sounds labeled contentment (14.8%) or sensual pleasure (11.7%). In addition, disgust sounds were mistaken as surprise (18.3%) and sounds of sensual pleasure were labeled sadness (17.4%).

Comparing recognition of non-verbal and verbal vocal expressions of emotions

In order to compare the recognition of emotion in verbal and non-verbal vocal stimuli, kappa scores were calculated for Experiments 2, 3, and 10. Note that happiness was not included in this comparison as it was not present in Experiment 2 and 3. Participants' performance with verbal stimuli was compared first with performance for non-verbal stimuli in Experiment 2, in order to include the maximum number of emotion categories. As the stimuli in Experiment 10 were selected as the best recognised stimuli from pilot testing, participants' performance in this task was then compared with the performance for the best non-verbal stimuli (Experiment 3). The mean kappa values for all emotion conditions in Experiment 2, 3, and 10 are shown in Appendix K.

Recognition performance with verbal and non-verbal stimuli

The overall kappa scores were tested using an ANOVA with emotion as a within-subjects factor and stimulus type as a between-subjects factor. There was a significant main effect of emotion ($F_{(8,369)} = 8.1$, $p < 0.0001$), indicating that participants were better at recognising some emotions than others (see Figure 5.1). There was no main effect of stimulus type, reflecting the fact that overall performance for the two stimulus types was very similar (mean kappa values 0.70 for speech stimuli and 0.67 for non-verbal stimuli). There was a significant interaction between stimulus type and emotion ($F_{(9,369)} = 20.1$, $p < 0.0001$), which is illustrated in Figure 5.1.

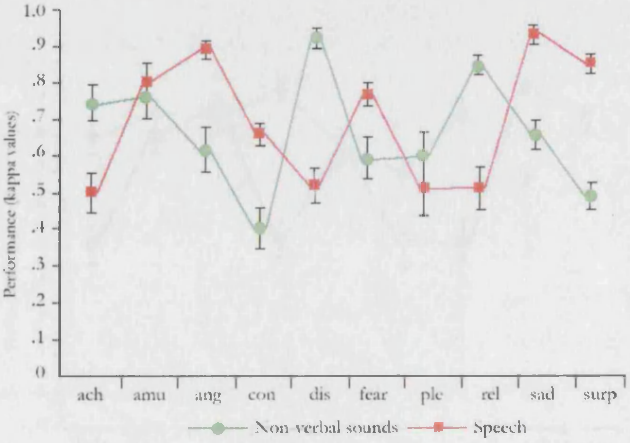

Fig. 5.1: Mean kappa values indicating recognition scores, per emotion condition for speech stimuli (Experiment 10) and non-verbal vocalisations (Experiment 2). Note: Ach = achievement/triumph, Amu = amusement, Ang = anger, Con = contentment, Dis = disgust, Ple = sensual pleasure, Rel = relief, Sad = sadness, Surp = surprise.

A series of independent samples t-tests were carried out to compare the recognition performance in Experiments 2 and 10 for each emotion. Bonferroni corrections were employed to control for multiple comparisons. Non-verbal stimuli were significantly better recognised than speech stimuli for expressions of achievement/triumph ($t_{(41)} = 3.0$, $p < 0.05$), disgust ($t_{(41)} = 7.2$, $p < 0.001$), and relief ($t_{(41)} = 5.0$, $p < 0.001$). Participants performed significantly better with verbal as compared to non-verbal stimuli, for anger ($t_{(41)} = 4.2$, $p < 0.001$), contentment ($t_{(41)} = 4.3$, $p < 0.001$), sadness ($t_{(41)} = 5.8$, $p < 0.001$), and surprise ($t_{(41)} = 7.2$, $p < 0.001$). There was also a strong trend for fear expressions to be better recognised from non-verbal stimuli ($t_{(41)} = 2.9$, $p < 0.07$).

There was no difference between the stimulus types for expressions of amusement and sensual pleasure.

Comparing performance with the best verbal and non-verbal stimuli

The kappa scores from Experiments 3 and 10 were compared using an ANOVA with emotion as a within-subjects factor and stimulus type as a between-subjects factor. There was a significant main effect of emotion ($F_{(8,328)} = 10.3$, $p < 0.0001$), indicating that participants were better at recognising some stimulus types than others (see Figure 5.2). There was also a main effect of stimulus type, reflecting the fact that participants were better at recognising non-verbal than verbal stimuli ($F_{(1,41)} = 17.5$, $p < 0.0001$; mean kappa values 0.70 for speech stimuli and 0.82 for non-verbal stimuli). There was a significant interaction between stimulus type and emotion ($F_{(8,328)} = 13.3$, $p < 0.0001$), which is illustrated in Figure 5.2.

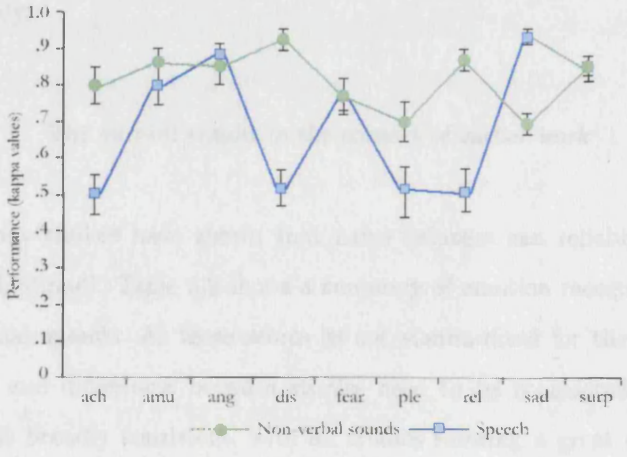

Fig. 5.2: Mean kappa values indicating recognition scores, per emotion condition for best speech stimuli (Experiment 10) and best non-verbal vocalisations (Experiment 3). Note: Ach = achievement/triumph, Amu = amusement, Ang = anger, Con = contentment, Dis = disgust, Ple = sensual pleasure, Rel = relief, Sad = sadness, Surp = surprise.

A series of independent samples t-tests were carried out to compare the recognition performance for each emotion with verbal and non-verbal stimuli. Participants recognised stimuli with higher accuracy for non-verbal as compared to verbal sounds of achievement/triumph ($t_{(}41) = 3.8$, $p < 0.01$), disgust ($t_{(}41) = 7.7$, $p < 0.001$), and relief ($t_{(}41) = 5.4$, $p < 0.001$). Participants performed significantly better with verbal as compared to non-verbal stimuli for expressions of sadness ($t_{(}41) = 6.4$, $p < 0.001$). There was no difference between the stimulus types for expressions of amusement, anger, fear, sensual pleasure, and surprise.

Discussion

The results of this study show that participants can reliably identify emotional expressions in speech. Listeners performed significantly better than chance at identifying the intended emotion for all stimulus types. For all stimulus types except for stimuli intended to express "happiness" the appropriate label was the most commonly selected one. Accuracy was by far the lowest for happiness sounds, with common confusions between happiness and some of the other positive emotion labels, especially contentment. This would seem to support Ekman's (1992b) hypothesis that the vocal emotion category "happiness" is better characterised as several positive emotions. However, it cannot be ruled out that the stimuli intended to express the broader category happiness were simply worse than the other stimuli. The design of this study attempted to minimise any obvious causes of stimulus differences, in using of the same speakers for all stimulus types and a consistent recording procedure. Nevertheless, it is impossible to rule out the possibility that sounds of "happiness" could be created that are as reliably recognised as vocalisations of the other emotions in this study.

The current results in the context of earlier work

A number of previous studies have shown that naïve listeners can reliably identify emotional expressions in speech stimuli. Table 5.2 shows a summary of emotion recognition in a number of studies with emotional speech. As these scores are not standardised for the number of response options, similarities and differences between studies need to be considered with caution. The overall pattern seems broadly consistent, with all studies showing a great deal of variability in the recognition levels of individual emotions, although recognition for all emotions is better than chance. Sounds of anger, sadness, and fear are better recognised, whereas listeners perform less well with disgust and happiness sounds. An exception is the study by Scott et al. (1997), where performance for happy stimuli was unusually high. That level of recognition is comparable to the recognition accuracy found in the current study for amusement sounds. Finally, it is also worth noting that recognition for surprise was high in the current study, although it has tended to be excluded in previous studies.

Tab. 5.2: Recognition rates (proportions) in studies of emotional speech. Options refers to the number of alternatives in forced choice tasks.

Study	Anger	Disgust	Joy	Fear	Sadness	Surprise	Options
Experiment 10[1]	.90	.57	.27	.79	.94	.87	11
Banse & Scherer, 1996[2]	.70	.15	.47	.59	.64	N/A	14
Juslin & Laukka, 2001	.58	.40	.51	.60	.63	N/A	6
Scherer et al., 2001	.79	N/A	.48	.74	.80	N/A	5
Scherer, 2003[3]	.77	.31	.57	.61	.71	N/A	N/A
Scott et al., 1997	.82	.59	.78	.49	.85	N/A	5

[1] Joy/Happiness based only on the category "happiness".
[2] Proportions correct calculated as average of cold anger/hot anger (anger), elation/happiness (happiness), panic fear/anxiety (fear) and despair/sadness (sadness). Confusions between cold anger/hot anger, elation/happiness, panic fear/anxiety or despair/sadness scored as correct.
[3] Based on a review of data.

Comparing recognition of verbal and non-verbal vocalisations

The close matching of the production of verbal and non-verbal stimuli (from Experiment 10 and 2, respectively) provided an opportunity to compare the accuracy for the two stimulus types. This analysis indicated that when non-verbal stimuli matched for recognition levels were used, the two stimulus types showed no overall difference in difficulty in emotion recognition. However, participants were significantly better at recognising non-verbal sounds when comparing the best available stimuli (Experiment 3) from the two stimulus sets. This indicates that non-verbal signals may be a more efficient means of communicating emotions than speech, although it is possible that this could be specific to these stimulus sets. Supporting this notion of non-verbal superiority is a study by Scott et al. (1997), which included verbal and non-verbal emotional vocalisations. Although no statistical comparison was made, the recognition levels were numerically higher for the non-verbal stimuli.

Most emotions showed differential recognition performance across the stimulus types. In the comparison with the non-verbal stimuli from Experiment 2, where stimuli were matched for recognition levels across the categories in that stimuli set, expressions of anger, contentment, fear, sadness, and surprise were better recognised from verbal stimuli, whereas achievement/triumph, disgust, and relief were better recognised from non-verbal sounds. There was no difference between the stimulus types for expressions of amusement and sensual pleasure. However, as the stimuli in the current experiment were selected for best recognition, a more appropriate comparison would be between the recognition of these stimuli and those in Experiment 3, with non-verbal stimuli selected for best recognition. In this comparison between the best verbal and non-verbal stimuli, the pattern was more consistent, with most of the emotions recognised better from the non-verbal stimuli. Only one emotion, sadness, was recognised better from verbal stimuli.

These findings indicate that although listeners can infer the emotional state of the speaker from both verbal and non-verbal sounds, some emotions are more accurately communicated in one type of vocalisation. Although most of the emotions were better recognised from non-verbal sounds, this was not universally the case. It is unclear whether there is one underlying determining factor or whether there might be a unique, possibly evolutionary, explanation for each emotion. Further work should explore how this pattern relates to the accuracy in identifying emotions in facial expressions. According to Ekman's original idea, the proposed positive emotions would have distinct, recognisable vocal expressions, but not unique facial expressions. He believed that all positive emotions would be communicated using one facial signal, the smile (Ekman, 1992b). To test that hypothesis is beyond the scope of this thesis, but the findings from this experiment indicate that some degree of specificity exists in the ease of communication using different types of signals.

Experiment 11 — Recognition of emotions in distorted speech

Twenty years ago, Scherer pointed out that reviews of the vocal communication of emotion "have revealed an apparent paradox: Whereas judges seem to be rather accurate in decoding meaning from vocal cues, researchers in psychoacoustics and psychophonetics have so far been unable to identify a set of vocal indicators that reliably differentiate a number of discrete emotions." (Scherer, 1986, pp. 143–144). This paradox remains despite the growth in research into vocal communication of emotions (Juslin & Scherer, 2005). A number of studies have now established that naïve listeners can infer a speaker's emotional state from vocalisations. However, little progress has been made in determining what acoustic cues are used in this communication. Studies that have reported acoustic cues for emotionally inflected speech have tended to include only a small set of features, mainly relating to pitch (see Juslin & Laukka, 2003). Pitch is thought to be a key feature for the communication of emotion in speech (Banse & Scherer, 1996; Bänziger & Scherer, 2005), although spectral detail is also believed to be important (Ladd et al., 1985; Murray & Arnott, 1993).

A theory of acoustic cues in emotional speech

The leading theory in the research of emotion in the voice is Scherer's componential model of emotion, which makes a number of predictions regarding the acoustic patterns of emotionally inflected speech (e.g., Scherer, 1986; 2001; 2003). As discussed previously, this thesis will not attempt to test the predictions from Scherer's model, as the main focus of this thesis is on non-verbal vocalisations of emotions. Testing the predictions of Scherer's model would certainly be too great an undertaking for this chapter, especially given the difficulty of measuring some of the specified acoustic parameters. The aim of this experiment is not to exhaustively examine the acoustic cues important for emotion in speech, but rather to examine the relative contribution of a number of acoustic features to emotion recognition in speech as compared to non-speech vocalisations.

Previous studies of acoustically manipulated emotional speech

The relative lack of research into the acoustic cues important for emotional communication in speech is likely due to the difficulty in measuring many of the acoustic cues that are thought to be important (Scherer, 1986). The current experiment therefore takes a different approach, using acoustic manipulations to disrupt pitch or spectral cues and testing emotion recognition using these manipulated stimuli. A few previous studies have used acoustically manipulated speech to study emotional communication. In an early study, Lieberman and Michaels (1962) used a forced-choice paradigm with speech synthesised to vary in F_0 and amplitude. The study included both categories such as boredom and doubt, and emotions, for example fear and happiness. Lieberman and Michaels found that when only pitch information was presented to the participants, the identification rate dropped to 44%, a drastic reduction from the 88% correct recognition found with the original sounds. When amplitude modulation alone was presented, performance dropped to only 14% correct, which although low, was still better than chance. The authors concluded that fundamental frequency, amplitude and fine structure all contribute to emotion recognition in speech.

A study by Ladd et al. (1985) used digital resynthesis to systematically vary intonation contour type, voice quality, and F_0 range of emotionally spoken speech segments. They hypothesised that the overall F_0 range and voice quality cues communicate the arousal state of the speaker, whereas the F_0 contour signals "cognitive attitudes" such as affective states. However, all three types of acoustic cues were found to influence listeners' judgments of both arousal and cognitive attitudes, leading Ladd et al. to suggest that the distinction between arousal and cognitive attitudes may not be mirrored in acoustic cues. Nevertheless, their findings did show that pitch and voice quality cues in speech affect listeners' perceived emotional state.

The current study

The manipulations used in the current study are the same as those used in Experiment 5, chosen because of the effects that they have on speech intelligibility. The manipulations are described in more detail in Experiment 5, but a brief summary follows here. Three different transformations were included: one-channel noise-vocoded speech, six-channel noise-vocoded speech, and spectrally rotated speech. One-channel noise-vocoding removes all the acoustic detail except the amplitude envelope of the sounds. Pitch and spectral structure are lacking and the speech cannot be understood (Shannon et al., 1995). Six-channel noise-vocoding removes most of the pitch and all the fine spectral structure, leaving the duration, rhythm and the broad spectral structure (Shannon et al., 1995). Six-channel noise-vocoded speech is intelligible after a brief training session (Scott et al., 2000). The third transformation was spectral rotation, which preserves amplitude envelope and duration information, pitch and pitch variation, while distorting both fine and broad spectral information (Blesser, 1972). Spectrally rotated speech cannot be understood due to the distortion of the broad spectral structure. Untransformed (but filtered) emotional speech was also included in the current experiment. For a summary of the acoustic cues in the four different conditions, see Table 3.1.

Method

Stimuli

Fifty-five stimuli (5 per emotion) were selected from Experiment 10, and low-pass filtered at 4 kHz. As was done with non-verbal stimuli in Experiment 5, copies of the selected stimuli were acoustically manipulated in three different ways in order to remove different acoustic types of information (see Table 3.1). The total stimulus set was 220 sounds, with 55 originals, 55 one-channel noise-vocoded sounds, 55 six-channel noise-vocoded stimuli, and 55 spectrally rotated sounds. These manipulations are described in more detail in Experiment 5. In each acoustic condition, there were 5 stimuli per condition expressing the following emotions: achievement/triumph, amusement, anger, contentment, disgust, fear, happiness, sensual pleasure, relief, sadness, and surprise. Examples of the stimuli are available in the additional materials.

Participants

Twenty British English speaking participants (9 male, mean age 25.7 years) were tested. The participants were recruited from the University College London Psychology department participant database. None had taken part in studies on non-verbal emotional vocalisations.

Design & Procedure

The participants performed an 11-way forced-choice task with the order of the stimuli randomised for each participant across emotional and acoustic manipulation conditions. The procedure and presentation was the same as the categorisation tasks in Experiment 10.

Results

Performance with manipulated speech stimuli

Chi-square analyses were carried out to test whether participants were categorising the stimuli at levels above chance (see Appendix L). Participants performed better than chance for all of the original sound categories except happiness. Performance varied greatly between emotions and manipulations (see Figure 5.3). In the one-channel noise-vocoded condition, sounds of anger, fear and contentment were recognised at above chance levels. In the six-channel noise-vocoded condition, amusement, anger, fear, relief, and sad stimuli were identified at better than chance levels. Finally, in the spectrally rotated condition, amusement, contentment, disgust and sadness stimuli were reliably recognised. No emotion type was reliably identified in all manipulations, and only happiness stimuli were not reliably recognised in any condition.

As was the case with the non-verbal stimuli, the use of the different response options differed greatly and so proportional scores were calculated, yielding a proportion of correct scores relative to the use of each response label in each condition. An ANOVA and planned comparisons were performed using these scores. This gave a main effect of emotion ($F_{(10,190)}$ =6.1, p < 0.0001), indicating that listeners were better at identifying some types of emotional vocalisations (see Figure 5.4).

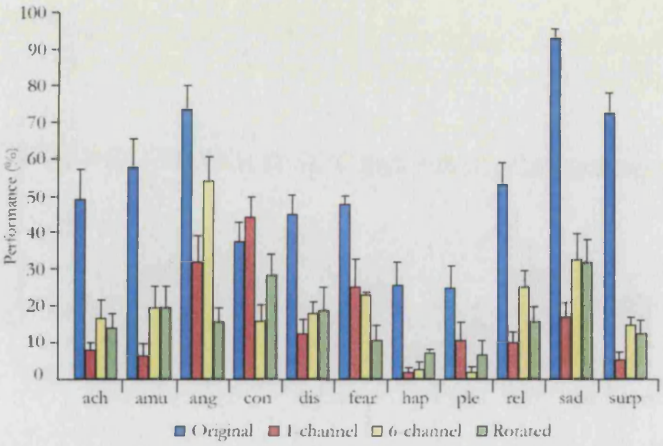

Fig. 5.3: Recognition of emotion in original and distorted speech sounds for each emotion (%). Note: Ach = achievement/triumph, Amu = amusement, Ang = anger, Con = contentment, Dis = disgust, Hap = Happiness, Ple = sensual pleasure, Rel = relief, Sad = sadness, Surp = surprise.

There was also a main effect of manipulation ($F_{(3,57)}$ =93.6, p < 0.0001). This pattern is illustrated in Figure 5.5. Three planned comparisons were carried out to explore this pattern in more detail. The first contrasted the original stimuli with all of the manipulated conditions, to test the prediction that the original stimuli would be the more accurately categorised. This contrast was significant ($t_{(19)}$=13.9, p < 0.0001) (mean score for original stimuli = 0.51, aggregate mean score for the manipulated conditions = 0.16). The second comparison contrasted the rotated and six-channel noise-vocoded stimuli, since for speech recognition, performance with six-channel noise-vocoded stimuli would be far better than with the rotated stimuli: this contrast was not significant (mean recognition for rotated stimuli = 0.17, mean recognition for the six-channel noise-vocoded stimuli = 0.19).

The third contrast was the one-channel versus six-channel noise-vocoded stimuli, since if the task was speech intelligibility, stimuli performance with the six-channel stimuli would be much better recognised than one-channel stimuli (Faulkner et al., 2001). This contrast was significant ($t_{(19)}$ =2.6, p = 0.016; mean recognition for one-channel noise-vocoded stimuli = 0.13). Finally, there was a significant emotion by manipulation interaction ($F_{(30,570)}$ =5.0, p < 0.0001). The pattern of this interaction was similar to that seen in Figure 5.3.

Comparing performance with speech and non-verbal stimuli controlling for response biases

In order to compare the recognition of emotion in non-verbal and verbal vocalisations of emotions, an ANOVA was carried out using the proportional accuracy scores from the current experiment

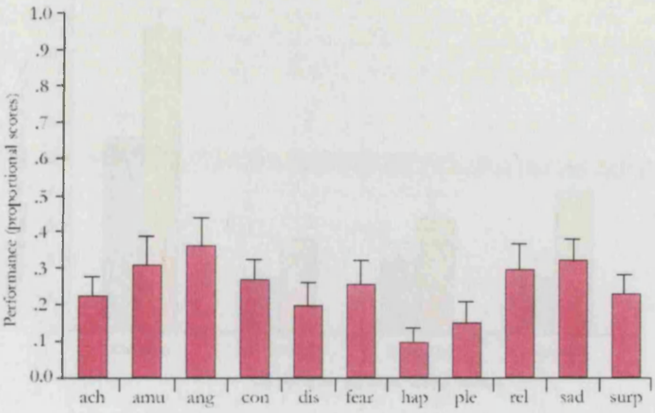

Fig. 5.4: Proportional recognition scores in emotional speech for each emotion (acoustic manipulations collapsed). Note: Ach = achievement/triumph, Amu = amusement, Ang = anger, Con = contentment, Dis = disgust, Ple = sensual pleasure, Rel = relief, Sad = sadness, Surp = surprise.

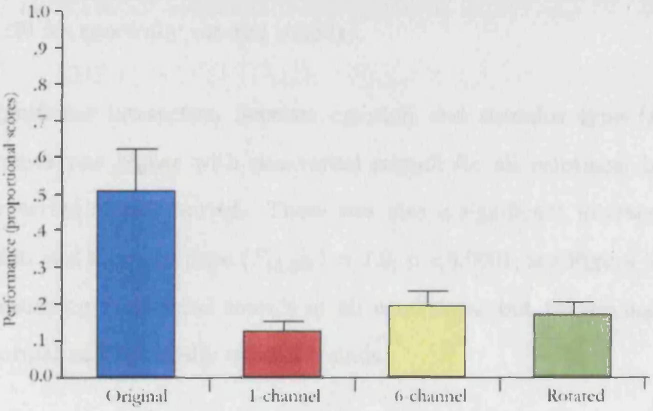

Fig. 5.5: Proportional recognition scores for acoustic manipulations of emotional speech (emotions collapsed).

and Experiment 5, with emotion and acoustic manipulation as within-subjects factor and stimulus type as a between-subjects factor. The category happiness was not included, as it was not present in Experiment 5. There was a main effect of group ($F_{(1,35)} = 21.2$, p <0.0001), with performance significantly higher for the non-verbal than the verbal stimuli (mean overall proportional accuracy scores .46 and .27, respectively).

There was also a main effect of emotion ($F_{(9,315)} = 8.9$, p <0.0001), reflecting the fact that participants recognised some emotions better than others across verbal and non-verbal stimuli. Performance ranged between .27 for sensual pleasure and .49 for amusement. There was also a main effect of acoustic manipulation ($F_{(3,105)} = 257.3$, p <0.0001), as participants were better

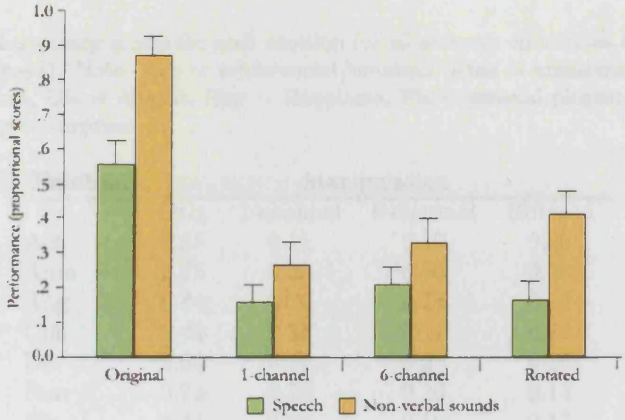

Fig. 5.6: Proportional accuracy scores for each acoustic condition with verbal and non-verbal stimuli (emotions collapsed).

able to identify emotions in some conditions compared to other conditions (average proportional accuracy scores .71 for original stimuli, .21 for one-channel noise-vocoded stimuli, .27 for six-channel noise-vocoded and .29 for spectrally rotated sounds).

There was a significant interaction between emotion and stimulus type ($F_{(9,315)} = 4.3$, p <0.0001). Performance was higher with non-verbal stimuli for all emotions, but the extent of the advantage over verbal stimuli varied. There was also a significant interaction between the acoustic manipulation and stimulus type ($F_{(3,105)} = 7.0$, p <0.0001; see Figure 5.6). Participants were better at categorising non-verbal sounds in all conditions, but to varying extents, with a stronger effect in normal and spectrally rotated sounds.

There was also a significant interaction between emotion and acoustic manipulation ($F_{(27,945)}$ = 5.3, p <0.0001; see Table 5.3). The original stimuli were best recognised for all emotions but the relationship between the other three conditions varied for different emotions. Finally, there was a significant three-way interaction between stimulus type, emotion and acoustic manipulation ($F_{(27,945)} = 1.9$, $p < 0.01$).

Comparing performance with verbal and non-verbal stimuli controlling for response options

As the analysis using proportional accuracy scores does not control for the different number of response options available in Experiments 3 versus 11, the analyses were repeated with kappa scores. Any calculations made with raw scores of zero resulted in negative kappa values. These

Tab. 5.3: Proportional accuracy scores for each emotion for all acoustic conditions (verbal and non-verbal stimuli collapsed). Note: Ach = achievement/triumph, Amu = amusement, Ang = anger, Con = contentment, Dis = disgust, Hap = Happiness, Ple = sensual pleasure, Rel = relief, Sad = sadness, Surp = surprise.

Emotion	Manipulation			
	Orig	1-channel	6-channel	Rotated
Ach	0.55	0.11	0.17	0.26
Amu	0.75	0.30	0.40	0.38
Ang	0.84	0.22	0.24	0.17
Con	0.41	0.35	0.11	0.24
Dis	0.66	0.11	0.21	0.25
Fear	0.74	0.10	0.20	0.14
Ple	0.44	0.16	0.10	0.15
Rel	0.62	0.13	0.29	0.28
Sad	0.73	0.15	0.32	0.34
Surp	0.67	0.12	0.28	0.26

were converted to a kappa of zero (385 responses in Experiment 11 and 196 in Experiment 3). An ANOVA was carried out, with emotion and acoustic manipulation as within-subjects factor and stimulus type as a between-subjects factor. There was a main effect of emotion ($F_{(9,315)}$ = 17.49, p <0.0001), reflecting the fact that participants recognised some emotions better than others. There was also a main effect of manipulation ($F_{(3,105)}$ = 12.73, p < 0.0001), as participants were better able to identify emotions under some acoustic conditions compared to others (average kappa values .63 for original stimuli, .17 for one-channel noise-vocoded stimuli, .21 for six-channel noise-vocoded and .22 for spectrally rotated sounds). There was a main effect of group ($F_{(1,35)}$ = 19,55, p < 0.0001), with performance significantly higher for the non-verbal than the verbal stimuli (mean overall kappa values .37 and .25, respectively). There were significant interactions between emotion and stimulus type ($F_{(9,315)}$ = 12,73, p < 0.0001), and between acoustic manipulation and group ($F_{(3,105)}$ = 16,92, p <0.0001), and between emotion and acoustic manipulation ($F_{(1,35)}$ = 7.98, p < 0.0001). There was also a significant three-way interaction between stimulus type, emotion and acoustic manipulation ($F_{(27,945)}$ = 4.43, p <0.0001).

Discussion

Spectral and pitch cues in emotional speech

This study investigated the role of acoustic cues in the recognition of emotions in speech. As expected, all of the acoustic manipulations caused a decrease in the listeners' accuracy in recognising emotions (see Figure 5.3). Although sounds in all manipulated conditions were less well recognised than the original stimuli there were also differences between the levels of recognition for the different manipulations.

Six-channel noise-vocoded stimuli were better recognised than one-channel noise-vocoded stimuli. This finding is in line with previous work on speech intelligibility showing that six-channel noise-vocoded speech is better understood than one-channel noise-vocoded speech, which is unintelligible (Faulkner et al., 2001). This indicates that, similarly to speech comprehension, the recognition of emotion in speech relies to some degree on broad spectral structure. This contrasts with the findings in Experiment 5 which showed that for non-verbal stimuli, performance was only marginally better for six- than one-channel noise-vocoded sounds. Together these findings suggest that broad spectral structure is more important for the recognition of emotion in speech sounds than non-speech vocalisations. As broad spectral structure is one of the key features used for understanding the content of spoken language, this stronger role for broad spectral structure for perceiving emotion in speech may be due to the interaction with intelligibility which is likely intrinsic to processing emotional speech.

In the current study, there was no difference between accuracy rates for spectrally rotated and six-channel noise-vocoded speech stimuli. This indicates that the participants were not utilising the pitch and pitch variation cues available in the spectrally rotated sounds.

Previous work on the acoustics of emotional speech

Previous work has suggested pitch to be a key feature in the perception of emotion in speech (e.g., Banse & Scherer, 1996; Bänziger & Scherer, 2005; Murray & Arnott, 1993). However, this work has generally been done using longer speech segments than the stimuli in the current study, and has tended to use acoustic analysis rather than selective removal of acoustic cues (see Juslin & Laukka, 2003).

Most studies that have discussed the role of pitch information in emotional speech have reported global measures of pitch, such as average F_0 or F_0 range (see Juslin & Laukka, 2003) although some more elaborate measures of F_0 have recently been proposed (Bänziger & Scherer, 2005). Together with duration and intensity, global measurements of pitch are relatively easy to measure. In contrast, spectral detail is more difficult to quantify, and although aspects of spectral detail are thought important for communicating emotions in speech, little work has studied this relationship empirically (Murray & Arnott, 1993). Scherer points out that

> "although fundamental frequency parameters (related to pitch) are undoubtedly important in the vocal expression of emotion, the key to the vocal differentiation of discrete emotions seems to be *voice quality*, that is, the timbre of the voice, acoustically determined by the pattern of energy distribution in the spectrum .... The reason for the neglect of voice quality in empirical studies of vocal affect indicators can be traced to the enormous conceptual and methodological difficulties encountered in the attempt to objectively define and measure these vocal characteristics." (Scherer, 1986, p. 145, italics in original).

It thus seems likely that the lack of research into the spectral characteristics of emotional speech is responsible for the minor role often attributed to these acoustic cues.

Acoustic cues for emotion recognition in speech and non-verbal sounds

For both verbal and non-verbal stimuli, all of the acoustically manipulated stimuli were less well recognised than the original sounds. This impairment can likely be attributed in part to the absence of accurate fine spectral detail in all of the manipulated conditions, an acoustic component that communicates cues such as voice quality. A previous study by Ladd et al. (1985) found that voice quality affected listeners' judgments of emotional speech. However, as Ladd et al. used rating scales rather than a forced choice task, the results of that study did not speak to the effect of voice quality on emotion recognition.

The pattern of the current study is different from the findings with non-verbal stimuli in Experiment 5, where spectrally rotated sounds were better recognised than six-channel noise-vocoded stimuli. This was interpreted as supporting the idea that pitch and pitch variation cues are important for the perception of emotion in human vocalisations. However, speech stimuli showed no improvement as pitch cues were added; that is, between the six-channel noise-vocoded stimuli and the spectrally rotated sounds. The interpretation of these findings is complicated by the fact that spectral and pitch information are not entirely independent. For example, six-channel noise-vocoded sounds contain both more spectral information and more pitch information than one-channel noise-vocoded sounds. Listeners performed better with the six-channel speech sounds, likely due to them utilising the additional broad spectral cues, although it may be partly due to them making use of the improved pitch. If it was entirely due to the use of pitch cues they would have been expected to also make increased use of the added pitch cues in the spectrally rotated sounds.

In sum, broad spectral cues are more important for emotion identification in speech, whereas pitch cues are more important for non-verbal sounds. The decrease in performance seen in all manipulated conditions as compared to the original sounds suggests that fine spectral detail plays an important role for recognition of emotions in both verbal and non-verbal vocalisations.

Recognition accuracy in speech and non-speech vocalisations

This study also provided an opportunity to compare recognition accuracy for acoustically manipulated speech and non-speech vocalisations in more detail. Overall, performance was significantly higher for the non-verbal than the verbal stimuli. This was reflected in all of the acoustic conditions, with participants consistently performing better with non-verbal stimuli. The extent of this advantage for non-verbal stimuli varied across acoustic conditions, with a stronger effect for normal and spectrally rotated sounds. This is in line with the finding in Experiment 10 which found that recognition accuracy was higher for non-verbal as compared to speech stimuli.

Across acoustic manipulations and stimulus types, participants recognised some emotions better than others. Sounds of amusement, anger and sadness were relatively well recognised, whereas contentment and sensual pleasure sounds were the least well recognised. Recognition accuracy for emotions also interacted with the stimulus type, such that performance was consistently higher for non-verbal stimuli for all emotions, but the extent of the advantage varied.

There was also an interaction between emotion and the acoustic manipulations, with the relationship between the acoustic manipulations varying for the different emotions. For example, for amusement the six-channel noise-vocoded sounds were better recognised than the spectrally rotated sounds. In contrast, disgust sounds were better recognised in the spectrally rotated manipulation than the six-channel noise-vocoded condition. Although the precise pattern for each emotion was not examined statistically, the interaction effect found between emotion and acoustic manipulation (illustrated in Figure 5.3) demonstrates that different cues are relatively more important for recognition of different emotions. Notably, the original, non-manipulated stimuli were best recognised for all emotions.

There was also a three-way interaction between emotion, stimulus type, and acoustic manipulation. This means that the acoustic manipulations differentially affected recognition for the different emotions, and that this pattern differed for verbal and non-verbal sounds. This complex pattern is difficult to interpret, especially without any specific hypotheses. Finally, it is worth noting that the same effects were found in analyses done with both kappa scores (controlling for the number of response options in the two experiments) and proportional scores (controlling for participants' response biases).

Experiment 12 — Intelligibility, speaker differentiation and emotion recognition in noise-vocoded speech

Different types of information in the human voice

Human vocalisations are rich signals, communicating not only the affective state of the speaker but also semantic information, and the speaker's gender, age, and social class (Karpf, 2006). Few attempts have been made to create a framework of how these different types of information are extracted and processed, and how these processes interact. Numerous theories exist regarding speech processing, but these have tended to neglect the paralinguistic aspects of the vocal signal.

Belin and colleagues recently proposed a model of voice perception, suggesting that the processing streams of different types of information about the voice are functionally separate (Belin et al., 2004). This model is a modified version of Bruce and Young's (1986) influential theory of face recognition. According to Bruce and Young's model, information about facial identity, facial expressions, and lip speech, are processed in separate, parallel routes. There is now extensive evidence from a range of disciplines supporting this model, although a modified version of it has recently been proposed (see Calder & Young, 2005). Belin et al. propose that voices are processed in a similar manner to faces, with information about the content of the speech, the identity and the emotional state of the speaker all being processed in functionally independent streams (see Figure 5.7).

In contrast to Bruce and Young's functional face processing model, Belin et al's voice processing account is "neurocognitive", that is, largely based on functional imaging data. The authors acknowledge that most imaging work studying the voice has investigated speech processing, and little is known about the processing of emotion and identity in the voice. According to Belin et al's model, speech is processed in anterior and posterior superior temporal sulcus (STS) and the

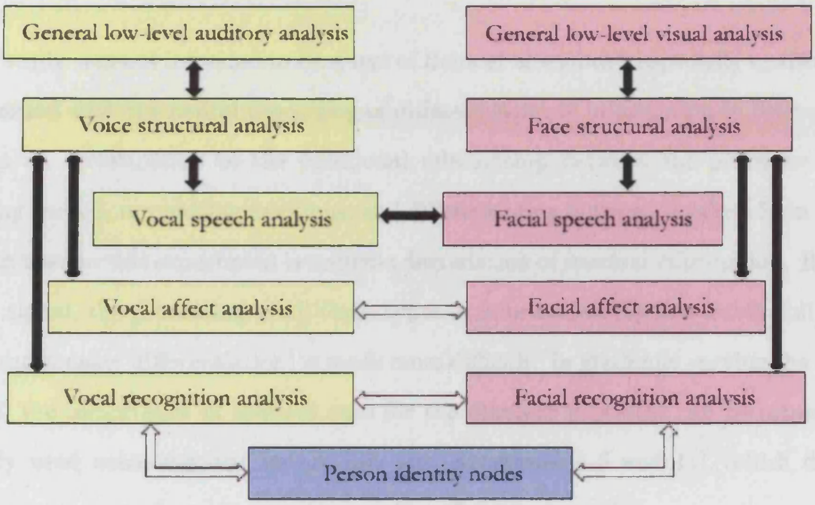
General low-level auditory analysis
General low-level visual analysis
Voice structural analysis
Face structural analysis
Vocal speech analysis
Facial speech analysis
Vocal affect analysis
Facial affect analysis
Vocal recognition analysis
Facial recognition analysis
Person identity nodes

Fig. 5.7: Belin et al's (2004) neurocognitive model of voice perception and its relation to face perception. Adapted from Belin et al. (2004)

inferior pre-frontal regions, predominantly on the left. A further distinction is made between the middle STS regions thought to be responsive to the mere presence of speech, and the anterior left STS/superior planum temporal plane, which Belin et al. propose to be involved in the comprehension of speech. Affective processing is hypothesised to take place mainly in the right temporal lobe, the right inferior prefrontal cortex, the amygdala and anterior insula. Speaker identity is suggested to be processed primarily in the right anterior STS. Given the lack of models of voice processing as a whole, Belin et al's model is a welcome attempt. However, it is difficult to assess with the currently available data: Although there is ample evidence for the localisation of speech processing to the left temporal lobe, only a handful of imaging studies have investigated emotion and identity in the voice (see Chapter 6 for a discussion of functional imaging studies of emotional vocalisations). Furthermore, studies investigating several of these processes within the same study are lacking (but see von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003). In addition, Belin et al. do not discuss any behavioral data supporting their model. This is somewhat surprising as behavioural dissociations would presumably be expected to underlie the neural dissociation they propose.

The approach in the current study

The current study was not intended to be a test of Belin et al's model, especially as their account is largely concerned with the neural processing of different types of information in the voice. Instead this study is an investigation of the functional relationship between the processes involved in understanding speech, recognising emotions, and differentiating between speakers from their voices. The approach used in this experiment is acoustic degradation of spectral information. By degrading the acoustic signal, the processing of all three types of information (speech intelligibility, emotion recognition and speaker differentiation) is made more difficult. In gradually varying the degradation of the signal, the importance of spectral cues for the different processes can be established. The current study used noise-vocoded speech (see also Experiments 5 and 11), which degrades the signal to varying extents depending on the number of channels used.

Perception of noise-vocoded speech

As discussed in Experiment 5, noise-vocoding preserves amplitude and temporal cues while removing spectral information. This is done by dividing the original signal into frequency bands, extracting the temporal envelopes from each band and using that amplitude envelope to modulate band-limited white noise. All bands are then summed back together (Shannon et al., 1995). Thus, the amount of spectral information in the signal is greater the more channels are used in the manipulation, allowing for a parametric investigation of the role of spectral cues in the processing of different types of information in speech.

Previous work with noise-vocoded speech has almost exclusively been focused on speech intelligibility. Research has established that naïve listeners can understand speech with three or more channels of noise-vocoding (Shannon et al., 1995), although the number of channels required depends to some extent on the difficulty of the material and the listening conditions (Shannon, Fu, & Galvin, 2004).

No work to date has looked at emotion perception in noise-vocoded speech, and little research has studied how noise-vocoding affects speaker identification. A recent study was the first to investigate speaker differentiation with noise-vocoded stimuli (Warren, Scott, Price & Griffiths, 2006). Listeners were asked to judge whether to speech segments vocoded with 1, 6, or 32 channels were produced by the same or different speakers. They found that the number of channels in the signal affected the participants' ability to accurately discriminate speaker identity, such that performance improved with an increase in the number of channels used. However, Warren et al. used only three levels of noise-vocoding and did not report extensive statistical analysis of the behavioural data. Nevertheless, it provides a demonstration that noise-vocoding affects listeners' accuracy in speaker discrimination as well as speech intelligibility.

Cochlear implants

Noise-vocoding is also a simulation of a cochlear implant (Shannon et al., 1995). A cochlear implant is a prosthetic hearing aid that converts sounds into electrical impulses that are delivered directly to the auditory nerve. Most patients with these implants show improvements with time in speech comprehension (Tyler and Summerfield, 1996). Research has also demonstrated that cochlear implant users can distinguish the gender of voices, although they may use different cues than normally hearing individuals (Fu, Chinchilla, Nogaki and Galvin, 2005). No research to date has investigated the perception of emotion in individuals with cochlear implants.

The aim of the current experiment

The aim of this experiment was to determine the importance of spectral cues for three types of voice processing: speech intelligibility, emotion recognition, and speaker differentiation. The acoustic structure of emotional speech stimuli was manipulated using noise-vocoding; Ten different levels were used, ranging from one to 32 channels, in order to parametrically vary the amount of spectral information in the signal. The effect of this manipulation was evaluated using one task for each of the relevant information types: speech intelligibility, emotion recognition, and speaker differentiation. The results from these three tasks were then compared to examine the extent to which these three processes rely on the same acoustic information.

Method

Stimuli

All of the 440 emotional speech stimuli recorded for the piloting stage of Experiment 10 were noise-vocoded at ten different levels: 1, 4, 6, 8, 12, 16, 20, 24, 28, and 32 channels. The original stimuli were not used in any of the tasks, that is, only noise-vocoded stimuli were included. The stimulus set included equal numbers of speech (three-digit numbers, see Experiment 10) of each of the eleven emotions achievement/triumph, amusement, anger, contentment, disgust, fear, happiness, sensual pleasure, relief, sadness, and surprise. Representative spectrograms are shown in Figure 5.8

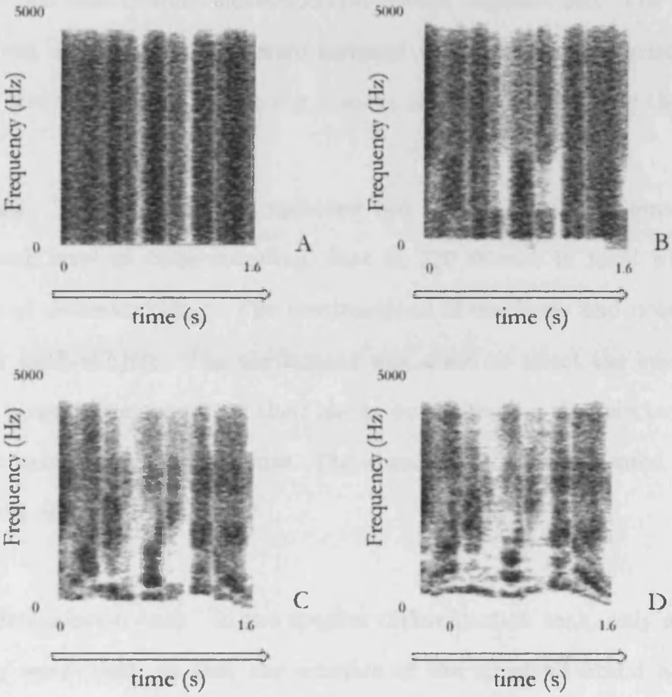

Fig. 5.8: Spectrograms of a speech stimulus expressing achievement/triumph at four levels of noise-vocoding: one channel (A), four channels (B), 12 channels (C), and 32 channels(D).

Participants

Twenty-nine British English speaking participants (10 male, mean age 25.5 years) took part in the experiment. The participants were recruited from the University College London Psychology department participant database. The participants had not taken part in any previous study with vocal expressions of emotions.

Procedure & Design

Each participant was tested individually in a computer cubicle. They completed the intelligibility task first, then the emotion recognition task and finally the speaker identity task. The order of stimuli within each task was random for each participant.

The intelligibility task This task consisted of ten trials, in which the participants heard a total of ten stimuli, one from each of the emotional categories except "happiness". In this task, each participant heard only one stimulus from each noise-vocoded level: one one-channel noise-vocoded speech segment, one four-channel noise-vocoded speech segment etc. The vocoding condition to which the different emotional stimuli were assigned was random and varied for each participant. The participant was required to report the content of the stimulus using the computer keyboard.

The emotion task The emotion task included two stimuli from each emotion category (except happiness) at each level of noise-vocoding, that is, 200 stimuli in total with 20 stimuli at each of the ten levels of noise-vocoding. The combinations of emotions and noise-vocoding conditions were random for each subject. The participant was asked to select the emotion expressed in the sound they just heard. They reported their choice by clicking on the selected response option on a computer screen using a computer mouse. The alternatives were presented distributed across the screen in the form of a circle.

The speaker differentiation task In the speaker differentiation task, only sounds from the "happiness" category were used, so that the emotion of the speakers would not confound listeners' judgments. In each of the 80 trials, the participant would hear two different stimuli. In half of the trials, the stimuli were spoken by the same speaker. In the remainder, where the stimuli were produced by different speakers, half of the trials were composed of stimuli spoken by two speakers of different genders. The participant's task was to report whether they thought the two stimuli were spoken by the same speaker or two different speakers. They chose by clicking on one of the options on the computer screen, using a computer mouse. The alternatives were presented horisontally on the screen.

Results

Performance on each task

As expected, participants' performance in the all three tasks improved with increased number of channels (see Figure 5.9). In all three tasks performance was around chance levels in the one-channel condition, but the range of improvement varied substantially between the tasks. In the intelligibility task, listeners performed at ceiling (97.8%) in the 32-channel condition. In the speaker differentiation task, listeners reached a maximum performance of 89.2%. Improvement was notably weaker in the emotion task, with performance never improving beyond 39%.

Logistic regressions were used to examine the relationship between the number of channels and participants' performance. Previous work has shown that models of the improvement seen in comprehensions of noise-vocoded speech with increasing channels is best fitted using a logarithmic transformation of the number of channels (Faulkner et al., 2001; Shannon et al., 2004). This is to account for the fact that the transition between one and two channels of noise vocoding has a greater impact than the transition between, for example, 22 and 23 channels. The statistical methods used in the current experiment followed this approach, using logistic regressions with a transformation using a logarithm with a base of 2 of the number of channels of the stimuli. The chi-square goodness-of-fit statistic of this analysis indicates how well the model fits the data. In this experiment, the chi test was only significant (i.e. the model differed significantly from the data, thus indicating a bad fit) for one participant in the speaker differentiation task, and for five participants in the intelligibility task. This suggests that the participants' performance in all three tasks was well modeled by the logistic regressions.

Logistic regressions were performed separately for the group data for each task. The model found an optimal solution with a slope of 0.71 for the intelligibility task, suggesting that participants' performance improved dramatically with increasing numbers of channels. In the emotion task, the logistic regression established a shallower slope of 0.43, indicating that, although emotion recognition improved with increasing numbers of channels, this improvement was much less dramatic. The improvement seen in the speaker differentiation task was intermediary to the other two tasks, with the logistic regression yielding an optimal solution at a slope of 0.61.

Emotion task — individual emotions

In order to examine participants' performance for each emotion in each channel condition, chi-square analyses were carried out using correct and incorrect scores (see Appendix M). However, these results are confounded by the participants' uneven use of the response options. For example, the anger label was heavily over-used for lower channel conditions, with 22% of the responses for one-channel stimuli and over 27% for four-channel stimuli being labeled anger (chance level = 10%). Thus, the chi-square analyses may not accurately represent the participants' performance

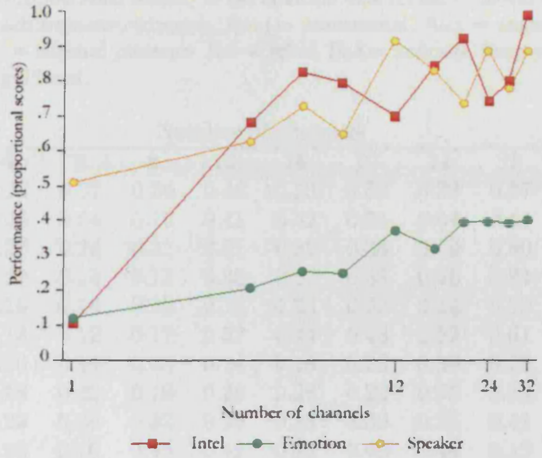

Fig. 5.9: Proportionate performance for all three tasks at each level of noise-vocoding. X-axis represents number of channels transformed using a logarithm with a base of 2. Note: Chance level is 10% for the intelligibility and emotion recognition tasks, and 50% for the speaker differentiation task.

on the task. To further analyse the data, proportional scores were calculated for each emotion for each channel condition (see Table 5.4). An ANOVA was carried out with the proportional scores, with emotion and channel condition as within-subjects factors. A main effect of emotion was found ($F_{(9,261)}$ = 16.2, p < 0.0001), as well as a main effect of channel condition ($F_{(9,261)}$ = 16.7, p < 0.0001) and an interaction ($F_{(81,2349)}$ = 1.9, p < 0.0001). This pattern shows that participants were better at recognising some emotions than others, that the number of channels affected the participants' performance, and that additional channels had a differential affect on performance for different emotions.

A series of ANOVAs were then performed, to examine whether recognition of each of the emotions was affected by the number of channels. The results were significant for all of the emotions except disgust and sensual pleasure (see Table 5.4). There was no observed relationship between participants' performance and the number of channels in the stimuli for disgust and sensual pleasure stimuli.

The speaker differentiation task

The participants' performance tended to respond "same" more often than "different", especially in the lower channel conditions. In fact, "same" responses comprised over 80% of responses for one-channel noise-vocoded sounds, and over 65% of responses for four-channel noise-vocoded stimuli.

Tab. 5.4: Performance (proportional scores) in the emotion task for each emotion at each channel condition. Note: Ach = achievement/triumph, Amu = amusement, Ang = anger, Con = contentment, Dis = disgust, Ple = sensual pleasure, Rel = relief, Sad = sadness, Surp = surprise, Ave = Average, N.S. = Not significant.

Emotion	Number of channels										F	P-level
	1	4	6	8	12	16	20	24	28	32		
Ach	0.10	0.29	0.37	0.34	0.46	0.33	0.52	0.39	0.37	0.45	6.22	0.000
Amu	0.07	0.45	0.64	0.10	0.41	0.42	0.54	0.54	0.44	0.49	6.87	0.000
Ang	0.17	0.22	0.24	0.32	0.51	0.32	0.38	0.69	0.60	0.53	3.93	0.000
Con	0.15	0.29	0.13	0.13	0.30	0.17	0.33	0.26	0.34	0.24	2.65	0.006
Dis	0.07	0.15	0.16	0.20	0.18	0.24	0.39	0.14	0.35	0.22	1.53	N.S
Fear	0.09	0.13	0.12	0.17	0.57	0.44	0.43	0.52	0.61	0.59	3.13	0.001
Ple	0.05	0.20	0.19	0.03	0.14	0.18	0.23	0.29	0.23	0.21	1.89	N.S
Rel	0.06	0.18	0.23	0.19	0.29	0.28	0.22	0.35	0.28	0.35	2.81	0.004
Sad	0.14	0.23	0.36	0.32	0.39	0.38	0.30	0.39	0.41	0.35	6.82	0.000
Surp	0.17	0.28	0.40	0.35	0.42	0.34	0.48	0.44	0.42	0.42	5.22	0.000
Average	0.11	0.24	0.28	0.22	0.37	0.31	0.38	0.40	0.41	0.38		

The proportion of "same responses" was around 50% for conditions with six or more channels of noise-vocoded stimuli. The data in Figure 5.10 is displayed using d' scores, which control for response biases.

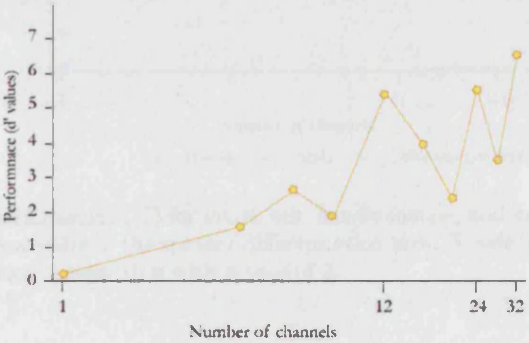

Fig. 5.10: Performance (d' scores) in each channel condition in the speaker differentiation task. X-axis represents number of channels transformed using a logarithm with a base of 2. Note: Zero is chance performance; higher d' value denotes better performance.

Given the substantial differences in F_0 between men and women's voices, (Fu et al., 2005) participants' scores were also calculated separately for stimuli with male-male, female-female and male-female stimulus pairs for both same-speaker stimulus pars and different-speaker stimulus pairs. For same-gender trials, participants' performance did not systematically shift across the channel conditions. This lack of an improvement in performance is likely due to participants' response bias in the lower channel conditions: Participants responded "same" to disproportionately many stimuli in the lower channel conditions and thus were accurate for the same-speaker stimulus pairs.

Participants' performance with different-speaker stimulus pairs is shown in Figure 5.11. In contrast to the same-speaker stimulus pairs, performance did improve with increasing numbers of channels. The participants' performance for all three types of stimulus pairs with different speakers (mixed sex, female-female, and male-male) thus show a similar pattern: Performance improved according to a logarithmic pattern, with added channels adding less and less to the participants' performance.

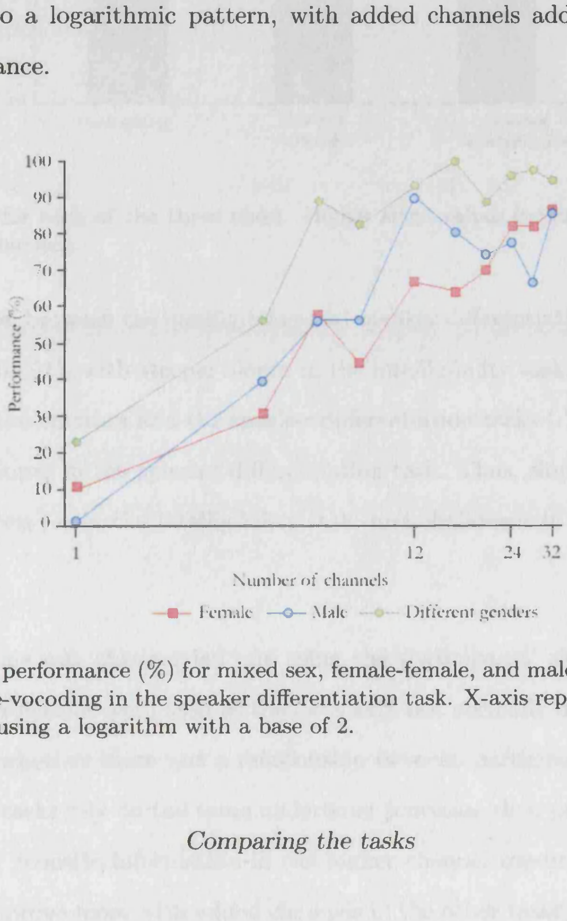

Fig. 5.11: Participants' performance (%) for mixed sex, female-female, and male-male stimulus pairs at all levels of noise-vocoding in the speaker differentiation task. X-axis represents number of channels transformed using a logarithm with a base of 2.

Comparing the tasks

In order to compare the three tasks, the slope of the improvement with additional channels was calculated for each participant for each task, using logistic regressions with the number of channels transformed using a logarithm with a base of 2. A higher slope value indicates a stronger improvement with added channels. An ANOVA was carried out using the slope-values, with task as a within-subjects factor, and found a significant effect ($F_{(2,56)} = 7.9$, $p < 0.001$). This pattern is shown in Figure 5.12.

Pair-wise comparisons were carried out, employing Bonferroni corrections to control for multiple comparisons. There was a significant difference between the intelligibility task and the emotion task (.419 mean difference, $p < 0.01$), with slopes being significantly steeper in the intelligibility

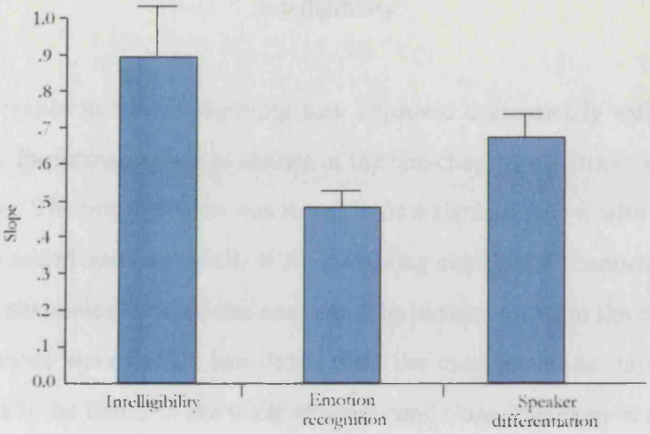

Fig. 5.12: Mean slopes for each of the three tasks. Higher slope values indicate a greater improvement with added channels.

task. The comparison between the intelligibility and speaker differentiation revealed a trend (.231 mean difference, p < 0.091), with steeper slopes in the intelligibility task. A trend was also found in the comparison of the emotion and the speaker differentiation tasks (.187 mean difference, p < 0.065), with steeper slopes in the speaker differentiation task. Thus, slopes were steepest — that is, improvement greatest — in the intelligibility task, and shallowest in the emotion recognition task.

A set of correlations was also carried out using the participants' slope values for the three tasks. Spearman's correlations were used as the data were not normally distributed. This analysis was used to examine whether there was a relationship between participants' performance on the different tasks. If the tasks rely on the same underlying processes then participants who are good at utilising the added acoustic information in the higher channel conditions for one task, would also be expected to improve more with added channels in the other tasks. However, there were no significant correlations between participants' slope values for any of the tasks.

Discussion

The results of this study showed that listeners' performance improved with added channels for all three tasks. Before discussing the relationship between the different tasks, each task will be discussed in turn.

Intelligibility

Participants' performance in the intelligibility task improved dramatically with an increase in the number of channels. Performance was at chance in the one-channel condition, and at ceiling in the 32-channel condition. The improvement was found to fit a sigmoid curve, with great improvement in performance with added spectral detail. With increasing numbers of channels present, the added information of more channels mattered less and less. The pattern found in the current study is thus consistent with previous work, which has shown that the most dramatic improvement is speech comprehension tends to be found in the lower channel-conditions (Shannon et al., 1995). This has been interpreted as implying that broad spectral cues are crucial for the comprehension of speech (Shannon et al., 2004).

The emotion task

There was an increase in performance — albeit slighter than that found in the intelligibility task — with added channels in the emotion recognition task. Performance was at floor with one channel, but rose to 39% correct with 20–32 channel stimuli. This suggests that broad spectral cues are important — though not sufficient — for successful recognition of emotions in speech, with performance improving as spectral information is added, although not as dramatically as in the intelligibility task. It is likely that voice quality cues — lacking in all channel conditions — are necessary for performance to improve beyond that level. Noise vocoding disrupts these cues regardless of the number of channels.

Participants were better at recognising some emotions than others. Considering this difference in overall performance, this could imply a differential reliance on spectral cues for different emotions. It could also be a consequence of the differences in recognition accuracy found with the original speech stimuli. However, the relationship between recognition accuracy and channel conditions varied with emotion type: additional channels improved participants' performance more for some emotions than others. This indicates that spectral cues are more important for some emotions, whereas for others, for example disgust and sensual pleasure (see Table 5.4), performance was not significantly affected by the number of channels, suggesting a lesser importance for spectral cues in the recognition of these emotions. These emotions may instead be identified largely on the basis of voice quality cues at the very fine levels of spectral detail.

The speaker differentiation task

As in the other tasks, participants' performance in the speaker differentiation task improved with increasing spectral resolution. The improvement was less dramatic than that seen in the intelligibility task but more substantial than in the emotion recognition task. Performance was at chance with one-channel noise-vocoded stimuli and reached 89% with 32-channel stimuli. Participants showed a response bias, tending to respond "same" more often than "different". This bias was particularly strong in the lower channel conditions, with "same responses" comprising over 80% of responses for one-channel noise-vocoded sounds and over 65% of responses for four-channel condition, but making up around 50% for all conditions with six or more channels. This indicates that with little spectral resolution, listeners tend to judge two stimuli are perceptually similar. Future studies should investigate whether this is a more general phenomenon or whether it is limited to judgments of whether the two sounds were produced by two different speakers.

In this task, no effects were found of the genders of the speakers or whether stimulus pairs contained same-sex or mixed-sex stimuli. This may seem surprising, given the large differences between the F_0 of male and female voices, with F_0 for male voices being on average 45% lower than those of females (see Fu et al., 2005). However, as noise-vocoded stimuli contain only limited pitch information, differences in pitch between male and female speakers would have been of little use to listeners in the current task.

Comparing the tasks

Participants' performance improved with added channels in all three tasks, although at different rates. The slopes were used as a measure of improvement with added spectral detail. Slope scores from the three tasks were compared to examine the participants' improvement with added channels. This analysis showed that participants' improvement in the intelligibility task was significantly greater than that seen in the emotion task. The improvement in the speaker differentiation task was in-between those of the other two tasks, and although the comparisons between the speaker differentiation task and the other two tasks failed to reach significance, strong trends suggested that these differences were not negligible.

Together these data indicate that spectral detail is most important for intelligibility of speech, and least important for emotion recognition. It thus seems that although spectral cues are important for performing for all three tasks, the processes rely to different extents on this information.

For intelligibility, broad spectral detail is crucial, but after a certain level of spectral resolution is reached, additional spectral information aids performance very little. This is shown by a dramatic increase in performance between one and four channels; subsequent channel increases are accompanied by smaller improvements in performance. In the emotion recognition task, improvement is more gradual, suggesting that the broad spectral cues so important for speech comprehension do not play the same role in emotion identification. Performance remained low in the emotion task, throughout channel conditions, and never improved to the levels seen with non-manipulated stimuli (see Experiment 10). This may imply that the fine spectral cues absent in all the noise-vocoded stimuli used play an important role for emotion recognition. The slope in the speaker differentiation task suggests that the improvement in performance with added channels is stronger than in the emotion recognition task, but weaker than in the intelligibility task. Performance does not increase dramatically between one and four channels, indicating that the broad spectral cues alone do not provide information crucial to speaker identification. It thus seems that the three tasks are showing three distinct relationships between performance and the amount of spectral information in the signal.

The current data do not rule out the possibility that some of these differences are due to the use of different tasks. The speaker differentiation task was the most constrained, in that it used a two-way forced choice. In the emotion recognition task, participants were required to choose between ten response alternatives. In the intelligibility task, listeners reported what they heard, but as they had been informed that what they would hear was a three-digit number, their responses would have been constrained, although less strictly so than in the other tasks. However, as the least constrained task resulted in the strongest increase in performance with added channels, it seems unlikely that the differences between the tasks are simply due to the different task constraints.

Individual performance in the three tasks

At the level of individual participants, there was no correlation between the improvements seen for the different tasks. Although caution must be used in interpreting null results, this pattern could indicate at least two things: that the cues used for each of the tasks are substantially different from those used in the other tasks, or that the underlying processes are different, such that being able to utilise acoustic cues for one purpose does not aid in the use of those cues for a different task. However, the data do not distinguish between these possibilities and so the conclusions that can be drawn here are limited.

Implications for Belin et al.'s model of voice perception

Belin and colleagues recently proposed a model of voice perception, suggesting that information about speech content, speaker identity and the emotion of the speaker are processed in functionally independent streams (Belin et al., 2004). The results of the current study lend some support to the idea of independence of the processing of these different types of information in the voice: Although the performance of listeners improved with added spectral details in all the three tasks, the relationship between participants' performance and the number of channels was not consistent across the tasks. The results of the current study indicate that the same acoustic information is used in different ways and is differentially important, depending on the task the listener is performing. In real life listeners of course extract many kinds of information at the same time, but the results of the current study seem to imply that this may be due to several (simultaneous) processes rather than one integrated process. These processes would rely to different extents on different kinds of acoustic information, and may also be processed in parallel but independent (neural) pathways. For example, Belin et al. have suggested that intelligibility would rely more on the areas in the left temporal lobe whereas information about speaker identity and emotion would mainly be processed in the posterior and anterior right STS, respectively. To test this hypothesis is beyond the scope of this thesis, but it could provide an interesting extension of the current study.

The model of voice perception put forward by Belin et al. was a modification of Bruce and Young's model of face perception (Bruce & Young, 1986). This model has been the framework used for the vast majority of research on face perception in the last 20 years. However, Calder and Young (2005) recently proposed a reinterpretation of this framework, where the pathways processing different types of information in the face are seen as being more connected. Calder and Young suggest that the different types of processing would be better conceptualised as relying on image-based principal components, some more important for the processing of one type of information and others important for several processes. At the centre of this view are the differences in the physical properties that are important for the different processes. The different processes rely on largely separate sets of principal components, but the visual information is not explicitly routed into separate pathways for extraction of affective, identity and speech information. This was demonstrated in a study by Calder et al., where a statistical analysis of facial images extracted largely independent sets of principal components that coded emotion recognition and face identity (Calder, Burton, et al., 2001).

A future challenge will be to evaluate Calder and Young's model in the context of voice processing: Do different aspects of voice processing rely differentially on different kinds of acoustic information in the same way that different aspects of face perception rely on different visual cues? The results of the current study are consistent with such a framework, but represent only a small step towards exploring this issue.

In sum, the results from the current study show that intelligibility, emotion recognition and speaker differentiation in speech are differentially affected by the removal of spectral detail. Added channels of noise-vocoding enhanced performance in all three tasks, although the degree and rate of facilitation differed. This pattern is interpreted as consistent with Belin et al's (2004) model of voice processing.

General Discussion

Emotion recognition in speech

Consistent with previous work, Experiments 10 and 11 demonstrated that naïve listeners can reliably identify emotional expressions in speech. Nevertheless, recognition rates were lower for speech stimuli than for non-verbal stimuli (Experiment 3) consistent with a previous study by Scott et al. (1997). Accuracy was also especially low for happiness sounds, consistent with Ekman's (1992b) suggestions that the vocal emotion category 'happiness' is better characterised as several distinct positive emotions.

As with the non-verbal stimuli, there was great variation not only in the recognition of individual emotions, but also in how they were affected by the acoustic manipulations used in Experiment 11. In terms of overall performance, one-channel noise-vocoded stimuli were recognised more poorly than six-channel noise-vocoded speech, suggesting that the recognition of emotion in speech relies to some degree on broad spectral structure. However, there was no difference in recognition accuracy between six-channel noise-vocoded and spectrally rotated stimuli, indicating that pitch cues are less important. This pattern contrasts with that found with non-verbal stimuli, where emotion recognition was found to rely more on pitch cues and less on broad spectral detail. However, a reliance on fine spectral detail, which communicates voice quality, was found for recognition of emotion in both types of stimuli relies to some degree, as seen in the decrease in performance seen in all manipulated conditions as compared to the original sounds.

The role of spectral cues for intelligibility, emotion recognition and speaker differentiation

The amount of spectral detail in the signal was found to affect participants' performance for intelligibility, emotion recognition and speaker differentiation. Performance improved most dramatically with added information in the intelligibility task, especially between one and four channels of noise vocoding. This supports the idea that broad spectral detail is important for speech comprehension (Shannon et al., 1995). In the emotion task, the increase in performance with added spectral resolution was much weaker, with performance failing to improve beyond 40% correct responses. This implies that although broad spectral cues are important, voice quality cues are especially crucial for emotion recognition in speech. As in the other tasks, participants' performance in the speaker differentiation task improved with increasing spectral resolution in the stimuli. The improvement was less dramatic than for intelligibility but steeper than for the emotion recognition task. Together, these findings imply the relationships between performance and the amount of spectral information in the signal are different for the three tasks. This finding is consistent with a number of models that have suggested that the processing of different types of information in the voice is separate (Belin et al., 2004), although the strictness of this separation has been questioned (Calder & Young, 2005).

6. AN FMRI STUDY OF NON-VERBAL VOCALISATIONS OF EMOTIONS

This chapter begins with a review of research investigating the neural underpinnings of emotional vocalisations, and links between perception and action of social signals. A functional imaging study involving passive listening to non-verbal emotional vocalisations in a sparse scanning paradigm is presented. The results show that perception of non-verbal emotional vocalisations automatically and robustly engages bilateral higher-level auditory regions, as well as a network of pre-motor cortical regions. This pattern suggests that hearing emotional sounds activates a preparatory motor response in the listener, similar to that seen during speech perception (Wilson et al., 2004) and facial expressions of emotions (Carr et al., 2003; Leslie et al., 2004).

The cognitive neuroscience of emotional communication has tended to focus mainly on the processing of facial expressions of emotions. However, in recent years, the interest in vocal signals of emotion has grown, along with the sophistication of available neuroimaging techniques. The majority of research into how emotional vocalisations are processed in the brain has focused on emotional speech. The current discussion is limited to data obtained with Positron Emission Tomography (PET), functional Magnetic Resonance Imaging (fMRI), and the study of neuropsychological patients.

The lateralisation of emotional vocalisations

A central issue in the cognitive neuroscience of emotional vocalisations is to what extent the processing of emotional speech is right-lateralised (see Pell, 2006). Although many studies show bilateral activation during the perception of emotional speech, activation is often stronger on the right than the left (e.g., George et al., 1996), and some have found activation exclusively in the right hemisphere (e.g., Mitchell et al., 2003).

Some argue that the processing of emotional speech is right-lateralised because the right temporal region is involved in the analysis of slow acoustic variations, for example supra-segmental information in speech, such as prosody. In contrast, the left temporal region mainly processes rapid acoustic changes, such as phonemes, important for speech intelligibility (Poeppel et al., 2004; Zatorre, 2001).

Others have suggested lateralisation on the basis of the response tendency that the stimulus elicits from the perceiver (Davidson, 1998; Davidson, Abercombie, Nitschke & Putnam, 1999). Davidson (1992) distinguishes between emotions that elicit approach and withdrawal. According to Davison et al., emotions associated with withdrawal are processed in the right hemisphere, while emotions that elicit approach are processed in the left hemisphere. These properties relate to the valence dimension proposed by Russell (1980), in that positive emotions tend to elicit approach, and negative emotions tend to elicit withdrawal. The empirical support for Davidson's model is weak, and a substantial number of studies have failed to find any evidence of lateralisation of valence (Buchanan et al., 2000; George et al., 1996; Kotz et al., 2003; Wildgruber et al., 2002).

Some authors have argued for a less absolute distinction between left and right hemisphere processes. Pell (2006) notes that evidence points towards the right hemisphere being especially important in the processing of emotional speech. However, he suggests that "the right hemisphere's participation in emotional prosody constitutes a *relative* rather than an *absolute* dominance in this processing domain ... the bulk of the findings nonetheless argue that left and right cortical regions share the burden of extracting meaning from emotional tone in speech" (p.222, italics in original).

Neuropsychological patient studies and neuro-imaging investigations into the processing of emotional speech and pseudo-speech

The literature on the processing of emotional prosody in neuropsychological patients is too extensive to review here, given the focus of this thesis on non-verbal vocalisations of emotions (for a review see Pell, 2006), but a few recent studies warrant discussion. Given the tendency of most studies to use only one type of task, the studies by Pell (2006) and Adolphs et al., (2002) are worth considering in more detail, as they both included several types of tasks and extensive patient groups.

Adolphs et al. (2002) tested 66 patients with focal lesions to the right, left, or both hemispheres. They used four semantically neutral sentences, spoken with hot anger, sadness, happiness, fear and positive surprise. Participants carried out a forced-choice task, and rated the stimuli on a scale from 0 to 5 for each of the different emotions and for arousal. Adolphs et al. found that damage to right fronto-parietal cortices (involved in pre-motor, motor and somatosensory functions) was associated with lower recognition scores. The results also suggested some involvement of right anterior temporal lobe and left frontal operculum. The authors interpret this as supporting a simulation theory of the perception of social signals. This issue is discussed in more detail below.

Pell (2006) tested 20 patients with focal lesions to the right or left hemisphere, using three types of tasks with pseudo-speech. He used an emotion discrimination task where listeners made same-different judgements, a forced-choice emotion identification task, and emotional rating tasks, where participants rated emotional sounds on a scale from 0–5, reflecting how much of the target emotion they thought the sound expressed. Both patient groups in Pell's study showed impaired comprehension of prosody, but there was some dissociation in the types of problems they displayed. Right hemisphere damage was associated with insensitivity to emotive features, affecting mainly the rating tasks, whereas left hemisphere patients had difficulty interpreting the prosodic cues in a language context, which affected their performance on the other two tasks. Pell concludes that understanding emotional prosody in speech recruits both left and right hemisphere regions, although the roles likely differ for the regions in the two hemispheres. Pell suggests that the right cortical regions are preferentially involved in extracting pitch variations, and mapping these patterns onto stored knowledge of their emotional meaning. Regions in the left hemisphere, in contrast, are "enhanced during emotion processing when emotive attributes of vocal signals are recognized as codes tied to, albeit not *within*, the language system, with mandatory reference to corresponding phonological and semantic structure" (p. 232, italics in original). This interpretation is consistent with the involvement of left temporal regions in the processing of emotional pseudo-speech (e.g., Grandjean et al., 2005), but it is not clear how this model would account for the involvement of left cortical regions during perception of non-verbal emotional vocalisations such as screams and laughs (Morris et al., 1999).

A number of functional imaging studies have shown that bilateral temporo-frontal areas are involved in the processing of emotional prosody. A PET study by Imaizumi et al. (1997) compared activation during the performance of a speaker identification task and an emotion recognition task, both with emotional word stimuli. Both tasks produced activation in the temporal poles bilaterally, while the emotion task also activated the cerebellum and regions in the frontal lobe. In a more recent study by Kotz et al. (2003), 12 participants performed a forced-choice task with sentences spoken in a happy, angry or neutral tone of voice. Processing of positive and negative intonation, as compared to neutral, was associated with an increase in bilateral fronto-temporal and subcortical (caudate) activation. These findings are consistent with a study by Wildgruber et al. (2002). Using four sentences, spoken in a happy, sad, or neutral tone, 12 participants were asked to identify the emotion and to judge which of two exemplars was more expressive. Emotional sentences were found to activate bilateral temporo-frontal cortices, as well as the anterior insula, thalamus, and pallidum. Bilateral activation has also been found using pseudo-speech. In a study by Grandjean et al. (2005), 15 participants listened to pseudo-speech with angry or neutral intonation. They found enhanced response during emotional as compared to neutral prosody in right and left STS.

Some studies have yielded right lateralised patterns of activation during the perception of emotional speech. In a study by Mitchell et al. (2003), 13 male participants listened to sentences spoken with happy, sad or neutral intonation. The contrast between emotional and neutral prosody showed increased activation in the right-lateral temporal lobe, specifically the middle and superior temporal gyri (STG). Buchanan et al. (2000) also found right-lateralised activation during the perception of emotional prosody. They compared activation during the performance of two different tasks. Ten male participants performed a target detection task with words, in the emotion condition on the basis of emotional tone, and in the verbal condition on the basis of the initial consonant sound of the word. Buchanan et al. found more activation in the right hemisphere in general, and in the anterior auditory cortex in particular, during the emotion task. The verbal task activated the left inferior frontal lobe. Similarly to Pell (2006), the authors conclude that regions in both hemispheres are involved in the perception of emotional speech, although the two hemispheres are preferentially involved in different aspects of the processing.

A study using a similar design with a more extensive range of emotional states was carried out by Wildgruber et al. (2005). Ten participants carried out an emotion identification task and a vowel identification task with semantically neutral sentences spoken in a happy, angry, fearful, sad or disgusted tone. Increased activation during the emotional task was found in the right frontal and temporal cortex, specifically the STS. Wildgruber et al. concluded that "the right auditory association cortex predominantly contributes to distinguish the emotional state of the speaker" (p. 1239).

The majority of studies to date have found bilateral activation in temporal regions during the perception of emotional prosody. Most of these studies have used the typical functional imaging paradigm in this area, where the processing of neutral speech is compared to that of emotionally inflected speech or pseudo-speech (e.g., Grandjean et al., 2005; Kotz et al., 2003; Mitchell et al., 2003). One exception to this is the relatively early PET study by Imaizumi et al. (1997), where participants performed a speaker identification task and an emotion task with the emotional word stimuli. In terms of fMRI studies, the study by Wildgruber et al. (2002) did include a task during scanning, but the contrasts relevant to this discussion compared listening to emotional over neutral speech, regardless of task. These studies differ from the studies that have found a right-lateralised response, which have compared activation during the performance of two different tasks, one emotional and one language task. Buchanan et al. (2000) used a target detection task with words, where participants would judge the emotional tone in one condition, and the initial consonant sound of the word in the other. Wildgruber et al. (2005) contrasted an emotion identification task and a vowel identification task. The use of these kinds of tasks in these studies is unlikely to provide a full explanation of the discrepancy in findings between them and the other studies. Nevertheless, it is worth noting that contrasting verbal to emotional tasks may result in different patterns of activations to comparing emotional to neutral speech.

The use of non-verbal vocalisations with neuropsychological patients

Although most of the work on emotion in the voice has studied emotional speech, a number of studies have also used non-verbal vocal stimuli with neuropsychological patients and neuroimaging. Most of the patient studies have investigated individuals with lesions to areas thought to be crucial to the processing of particular emotions (see below), but a couple a studies have focused on patients with lesions to the frontal lobes. A study by Hornak, Rolls and Wade (1996) investigated patients with frontal lobe damage. Hornak et al. used the Ekman and Friesen face stimuli (1976) and non-verbal vocalisations expressing sadness, anger, fear, disgust, contentment, puzzlement, and neutral affect (brief monotone humming sounds). The patients were given forced-choice tasks as well as a series of questionnaires assessing behavioural changes. The results showed that patients with ventral frontal lobe damage were impaired in their recognition of facial and vocal signals of emotions, and the extent of this impairment was correlated with behavioural problems. In a follow-up study by the same group of researchers, emotion recognition was examined in patients with damage to different pre-frontal regions (Hornak et al., 2003). This study used a similar methodology, consisting of non-verbal vocal stimuli, pictures of emotional facial expressions, and emotional change questionnaires, although the facial expressions stimuli were morphed images including a blend of two emotions, rather than the original Ekman and Friesen set of prototype expressions. The results revealed that patients with unilateral orbital lesions or medial lesions to BA 9 or the anterior ventral Anterior Cingulate Cortex (ACC) were impaired in the recognition of non-verbal vocal signals of emotions. In contrast, patients with dorsolateral lesions or medial lesions outside of BA 9 and the anterior ventral ACC were not impaired, compared to controls. None of the groups were consistently impaired in the recognition of facial expressions, a somewhat surprising finding, given the involvement of pre-frontal regions in the processing in emotional faces (e.g., Iidaka et al., 2001).

Several studies have used non-verbal vocalisations to investigate patients with amygdala lesions. The amygdala has been implicated in the processing of signals communicating fear (Calder, Lawrence, & Young, 2001), and these patients are thus a candidate group for exhibiting selective impairments to the processing of a single emotion, fear. The first study to examine the perception of emotion with non-verbal vocal stimuli in a patient with bilateral amygdala lesion, was a study by Scott et al. (1997), which reported the case of patient DR. This patient had previously been reported to be impaired in the recognition of fear signals from facial expressions (Calder, Young, Rowland, Perrett, Hodges & Etcoff, 1996; see also Adolphs et al., 1999 for a similar finding with a group of patients with amygdala lesions). Using both verbal and non-verbal emotional vocalisations, Scott et al. showed that DR was impaired at recognising fear and anger signals from both verbal and non-verbal vocalisations of emotions. This demonstration was the first to show that amygdala damage can cause emotion recognition impairments from other cues than facial expressions.

Scott et al.'s finding was replicated with a different patient by Sprengelmeyer et al. (1999). They tested another patient, NM, with bilateral amygdala damage, using facial expressions, non-verbal vocalisations, and body-posture stimuli expressing the basic emotions. Sprengelmeyer et al. found that NM was selectively impaired in the recognition of fear from all types of stimuli, supporting the notion of the amygdala playing a crucial role in the perception of fear signals in several modalities.

This conclusion has been questioned by a study by Anderson and Phelps (1998). They tested SP, a patient with bilateral amygdala damage, who unlike patient DR in Scott et al.'s study had intact basal ganglia. Anderson and Phelps argued that the impairments found by Scott et al. may have been due to the patient's damage to the basal ganglia, which has been shown to be important for prosody evaluation. SP had previously been shown to have an impairment in the recognition of facial expressions of emotions (Phelps, LaBar, Anderson, O'Connor, Fulbright, & Spencer, 1998). Anderson and Phelps tested SP with verbal and non-verbal emotional stimuli, and found that she was able to reliably recognise expressions of both fear and anger, although recognition was somewhat low compared to control participants for non-verbal stimuli. SP was dramatically impaired in the recognition of disgust from prosody, although she performed within the normal range for non-verbal disgust stimuli. Anderson and Phelps claim that these data show that amygdala damage may be necessary, but not sufficient for an impairment to emotional recognition from auditory cues. They conclude that "the analysis of nonverbal signals of fear from different input channels are dissociable, being at least partially dependent on different brain structures" (p. 3607). They point out that the amygdalae receive extensive input from visual, but not auditory cortices. They suggest that the amygdalae receive auditory information from striatal areas, and that these amygdalo-striatal interactions form the crucial aspect of emotion perception from auditory cues.

In addition to fear, disgust has been argued to be processed by a distinct neural circuit involving the insula and basal ganglia (Calder, Lawrence, et al., 2001), A study by Calder, Keane, Manes, Antoun & Young (2000) examined NK, a patient with a left hemisphere lesion to the posterior part of anterior insula, posterior insula, internal capsule, putamen and globus pallidus. They showed that NK was selectively impaired at recognising signals of disgust from facial, verbal and non-verbal emotional stimuli. Calder et al. also found that although NK's knowledge of the concept of disgust was unimpaired, he was less disgusted than controls by disgust-provoking scenarios.

Functional imaging work with non-verbal vocalisations

Four studies to date have used non-verbal vocalisations in functional imagining paradigms. While three of these investigated the neural processing of non-verbal vocalisations in the whole brain, one study specifically examined the effect of listening to laughter and crying on the amygdala (Sander & Scheich, 2001). In that study, 11 participants listened passively to blocks of laughing or crying sounds, produced by professional actors. The auditory baseline consisted of blocks in which participants performed a mental arithmetic task with heard digits. Sander and Scheich found that both laughing and crying activated the amygdala bilaterally, although laughter elicited a stronger response. In addition, they examined the effect of listening to laughter and crying on the auditory cortex. They found that this area was activated bilaterally by both laughter and crying, although activation was stronger for laughing than crying sounds, likely due to basic acoustic differences between the two types of auditory stimuli. Sander and Scheich also examined whether the perception of laugher and crying sounds would cause increased activation in the insula. They found that both types of sounds activated the insula bilaterally, consistent with the idea that this region is involved in the articulation of vocalisations (Dronkers, 1996). The authors conclude that the amygdala is involved in the processing of emotional stimuli from auditory as well as visual cues.

In a study using PET, Morris et al. (1999) investigated the neural processing of non-verbal vocalisations expressing happiness, sadness and fear in six male participants. Six male participants performed a gender decision task whilst listening to the emotional vocalisations and neutral sounds (humming noises). Listening to emotional over neutral stimuli was found to activate the left middle temporal gyrus, the left superior frontal gyrus, the right caudate nucleus, bilateral insula, and bilateral ventral pre-frontal cortex. The study also specifically examined patterns of activation during fear sounds compared to all other conditions. Morris et al. found that the perception of fear sounds increased activation in the left anterior insula, but decreased activation in the right amygdala and right insula. The authors point out that the left lateralised responses they found may have been due to the gender decisions task they used, which has been found to produce left-lateralised responses with facial stimuli (Morris et al., 1996). They suggest several possible explanations for the decrease in activation in the right amygdala found during fear sounds. It may have been due to fear-specific inhibitory interaction with the insula, or the rapid habituation of activation in the amygdala. Morris et al. conclude that emotional vocalisations are processed by a network including temporal and pre-frontal cortices, as well as the insula and amygdala nuclei.

This general interpretation is consistent with an fMRI study by Phillips et al. (1998), which investigated the perception of emotion from both facial and non-verbal vocal expressions in six male participants. The researchers used expressions of fear and disgust, as well as "neutral" expressions, with which participants performed a gender decision task. The facial stimuli were taken from the Ekman and Friesen set (1976), with the neutral stimuli being morphs consisting of 70% neutral and 30% happy expressions. The vocal stimuli were taken from the set used in the study by Scott et al. (1997), with the happy stimuli used as neutral expressions. Phillips et al. found that both kinds of fearful stimuli activated the amygdala, although the visual fear stimuli activated the left amygdala and the auditory fear stimuli activated the right amygdala. This finding is consistent with the localisation of fear processing in the amygdala (Calder, Young, Rowland, et al., 1996; Calder, Lawrence, et al., 2001). In terms of disgust stimuli, Phillips et al. found that visual disgust stimuli activated the anterior insula, the caudate nucleus, and putamen, none of which were activated by the disgust sounds. The authors interpret these data to mean that disgust signals in different modalities may be processed in different neural regions. All types of stimuli activated the STG, and all but the disgust sounds also activated the middle temporal gyrus (MTG). Phillips et al.

point out that the superior temporal sulcus, the sulcus between the superior and medial temporal gyri, has previously been shown to be involved in reading social signals from the face (Kanwisher, McDermott, & Chun, 1997) as well as in word perception (Fiez, Raichle, Balota, Tallal & Petersen, 1996). The authors interpret the activation seen in their study as reflecting the role of this region in decoding social signals.

Temporal regions were also implicated in a recent study by Meyer et al. (2005). In this study, 12 participants performed an auditory target detection task while listening to speech, laughter or non-vocal sounds (superimposed and connected single frequencies). Listening to speech compared to laughter activated left temporal regions, including the lateral STG and STS. Hearing laughter activated the right STG, and specifically activated regions involved in voluntary motor (larynx) functions in the right pre-central gyrus and sub-central gyrus. This may indicate that some of the brain regions implicated in the production of laughter are also involved in the perception of laughter sounds. Both types of vocal sounds (speech and laughter) compared to the non-vocal sounds activated bilateral sections of the anterior STS, with more extensive activation in the right hemisphere. This finding is consistent with the suggestion that the right STS is specifically involved in the processing of human voices. This claim is discussed in more detail below.

The STS — an area selectively tuned to the human voice?

In a study by Belin et al. (2000), the hypothesis of a voice-specific area was explored. In three experiments, participants listened to human vocalisations and a number of control conditions including energy-matched non-vocal sounds, human non-vocal sounds, scrambled voices, and white noise modulated with the same amplitude envelope as the vocal sounds. In all three experiments, peaks of voice selectivity could be found in most subjects along the upper bank of the STS bilaterally, with stronger activation in the right hemisphere. In a related study by Belin et al. (2002), eight participants listened to speech and non-verbal human vocalisations (e.g., coughs, screams). The control sounds consisted of environmental and musical sounds, and scrambled versions of the vocalisations. In Experiment 1, vocalisations were compared to non-vocal sounds, and activation was found in the STS bilaterally, both anterior and posterior to primary auditory cortex (Heschl's gyrus), with stronger activation on the right side. These peaks were used in Experiment 2, in which activation during the perception of speech sounds was compared to that during non-verbal vocalisations, within the regions identified as "voice selective" in Experiment 1. Speech sounds yielded stronger activation in most of the regions, including the primary auditory cortex. Speech and non-verbal vocalisations were also separately compared with scrambled versions of the same sounds. Speech sounds elicited greater activation than their scrambled counterparts in almost all of the specified regions, whereas the non-verbal vocalisations yielded a greater response than the scrambled non-verbal sounds specifically in the middle and anterior right STS. The authors interpret these data as support for the voice-selective area located in the anterior STS bilaterally, with the area on right being involved in paralinguistic processing of vocal stimuli. They point out that the left STS was more active during speech than non-speech vocalisations, consistent with previous work that has shown this area to be sensitive to speech intelligibility (Scott et al., 2000).

Several problems exist with Belin et al.'s (2000) model. A crucial test for the claim that this region is human voice specific would be to compare human voices with vocalisations from closely related species, such as apes or monkeys. Although such a study was recently carried out, a direct comparison between human and monkey vocalisations was not reported (Belin, Fecteau, Charest, Nicastro, Hauser & Armony, 2006). Another possible caveat with the proposed voice selective region is that it is heavily based on a model which has recently been seriously questioned. The idea of a voice specific area is proposed to be the auditory counterpart to the face selective area which has been claimed to exist in the fusiform area (Kanwisher et al., 1997; McCarthy, Puce, Gore, & Allison, 1997). This region has been suggested to be involved in processing visual stimuli of high expertise, rather than human faces per se (Gauthier, Skudlarski, Gore & Anderson, 2000). This may turn out to be the case for the proposed voice selective area as well. This hypothesis would be easily testable, by for example examining whether ornithologists exhibit increased activation in the proposed voice selective area while listening to bird song. An additional problem with Belin et al.'s account is the finding that the upper bank of the superior temporal sulcus, the specific region proposed to be voice specific, has been shown to be multi-modal, rather than auditory-specific, in monkeys (e.g., Barnes & Pandya, 1992). There is extensive work showing that much of the STS in humans is also multi-modal (Calvert, Campbell & Brammer, 2000; Calvert, Hansen, Iversen & Brammer, 2001; for a review see Calvert, 2001). Belin et al. point out that an early study of macaques showed that a region homologous to those identified in their study receives input exclusively from auditory-related areas (Seltzer & Pandya, 1978), and argue that the regions identified in their paper as voice selective may be part of a system that is involved in high-level analysis of complex acoustic information, and transmission of this information to other areas in the STS for multimodal integration.

Does perceiving emotional signals elicit preparatory motor action?

In addition to regions involved in higher-level processing of complex signals, such as the STS, some studies have found that social signals elicit activation in regions involved in the production of those same actions (Carr et al., 2003; Leslie et al., 2004). This link between perception and action has become known as the "mirror-neuron hypothesis". Mirror neurons, originally identified in area F5 in ventral pre-motor cortex in monkeys, fire both when an animal performs an action and when it observes the same action performed by another individual (Gallese, Fadiga, Fogassi & Rizzolatti, 1996). This finding has now been extensively replicated and has been found to extend to sounds (Kohler, Keysers, Umit, Fogassi, Gallese & Rizzolatti, 2002) and communicative gestures (Ferrari, Gallese, Rizzolatti & Fogassi, 2003).

Recently, evidence has been presented that suggests that mirror neurons exist in humans. In a study by Krolak-Salmon et al. (2006) an epileptic patient was implanted with depth electrodes in areas of the left frontal and temporal lobes, including the pre-supplementary motor area (pre-SMA). Stimulation of electrodes in this area was found to produce laughter and smiles in the patient, consistent with a previous study by Fried et al. (1998), which reported laughter triggered by electrical stimulation in the left pre-SMA. Krolak-Salmon et al. also demonstrated that ERP recordings from the same region in the patient showed selective activation during viewing of happy facial expressions, as compared to other emotional expressions. The authors point out that this region has extensive connections to the STS, and suggest that the left pre-SMA may play an important role in the link between perception and action, particularly for positive affect.

The mirror neuron hypothesis of social signals, sometimes referred to as "the motor theory of empathy" (Leslie et al., 2004), has also been investigated in humans using functional imaging. A study by Carr et al. (2003) explored the hypothesis that the understanding of others' emotional expressions is mediated by empathy using action representation, such that we understand others' emotional states by our brains simulating the associated action, which allows us to understand the other's affective state. They hypothesised that imitating an emotional expression and simply observing it, would activate similar regions, including the STS, the inferior frontal gyrus (IFG) equivalent to monkeys' area F5, and the anterior insula. According to their model, temporal regions including the STS are involved in higher level perceptual processing of the incoming stimulus, the IFG codes the goals of the perceived action and sends efferent copies of the motor plans back to the STS, to be compared to the perceived external stimulus. The insula links these regions with the limbic system, enabling an emotional evaluation. Although the model would predict that both imitation and perception would activate these regions, imitation would be expected to cause greater activation, as it involves additional feedback from actual movement. Carr et al. confirmed these hypotheses in an fMRI study, in which 11 participants viewed and imitated emotional facial expressions of the six basic emotions from the Ekman and Friesen (1976) set. The results showed that activation during the two tasks was largely overlapping, and that activation was stronger during imitation than perception. Activation was found in motor and pre-motor regions including the pre-SMA and the IFG, as well as in the STS, insula, and amygdala. The authors conclude that "we understand the feelings of others via a mechanism of action representation shaping emotional content, such that we ground our empathic resonance in the experience of our acting body and the emotions associated with specific movements" (p. 5502).

A similar finding comes from an fMRI study by Leslie et al. (2004), comparing imitation and passive viewing of finger movements and facial expressions in 15 healthy participants. Both types of imitation activated the left IFG, including Broca's area, as well as bilateral pre-motor areas, and bilateral SMA. In contrast to Carr et al.'s (2003) study, the left IFG was not activated during passive viewing of either faces or hands. Leslie et al. suggest that this may be due to the limitations of the 1.5 Tesla scanner they used, or that the stimulus set used may not have been adequate. They did find overlap between viewing and imitating conditions in the right ventral pre-motor area, although activation was more bilateral during the imitation task. The authors conclude that these findings support the motor theory of empathy, and that the right pre-motor cortex may be involved in both the generation and perception of social signals.

Although studies in humans have established that visual social signals elicit activation in areas involved in motor preparation, no studies have been done using vocal stimuli. In monkeys, auditory stimuli associated with actions have been found to active area F5 (Kohler et al., 2002), but this work has yet to extended to humans. One class of auditory social stimuli that has been shown to activate areas involved with motor functions in humans is speech. Several studies have established that listening to speech sounds activates regions involved in speech production, using both fMRI (Wilson et al., 2004) and transcranial magnetic stimulation (Fadiga et al., 2002; Watkins, Strafella, & Paus, 2003). However, this link between perception and action for other social auditory signals has not yet been investigated, and studies using both motor- and acoustic baselines are lacking.

Experiment 13 — The neural processing of non-verbal vocalisations of emotions

Hypotheses of the current study

This study aimed to investigate whether the link between perception and action previously found for speech and emotional facial expressions would also exist for non-verbal vocal expressions of emotions. Based on previous work, it was hypothesised that areas involved in motor planning, such as the IFG (Carr et al., 2003), and the left anterior insula (Dronkers, 1996) would be activated during the perception of vocal social signals. The pre-SMA was expected to be activated, particularly during the perception of positive emotions (Krolak-Salmon et al., 2006). On the basis of previous work using non-verbal vocalisations, it was hypothesised that secondary auditory regions along the temporal lobe would be activated (Meyer et al., 2005; Morris et al., 1999). In particular, the STS was expected to be activated, due to its proposed status as a voice selective area (Belin et al., 2000). Finally, based on research on emotional prosody in speech, the pattern of activation were expected to be right-lateralised (Buchanan et al., 2000; Wildgruber at al., 2005) or bilateral (Grandjean et al., 2005; Kotz et al., 2003).

Methods

Stimuli

The stimulus set used in this experiment consisted of non-verbal vocalisations of achievement/triumph, amusement, disgust, and fear. Twenty stimuli from each of the four emotion categories were selected from the body of stimuli recorded for Experiments 1 and 2. On the basis of pilot data, the 20 best recognised stimuli from each emotion category were selected. The acoustic baseline consisted of concatenated, spectrally rotated versions of the stimuli from the relevant emotion categories (see Experiment 5). To ensure that the spectrally rotated stimuli were not perceived as expressing any of the emotions included in the study, five participants (4 male, mean age 30.0 years) carried out a categorisation task with the spectrally rotated stimuli. These participants did not take part in the main study. In the categorisation task, listeners were played the 15 spectrally rotated tokens and asked to categorise them as achievement/triumph, amusement, disgust, fear, or "none of the above". Out of a total of 75 responses, 43 were "none of the above", 11 were amusement, 8 were disgust, 7 were achievement/triumph, and 6 were fear. There was no stimulus where a single emotion label was more common than the "none of the above" label, confirming that the sounds were not consistently perceived as expressing any single emotion.

Participants

Twenty right-handed participants took part in the experiment (8 males, mean age 32.5 years). All participants had normal hearing and no history of neurological problems. None of the participants had taken part in any previous experiments with emotional vocalisations and all gave informed consent to take part in the experiment.

Imaging procedure

Data were obtained using a Philips Intera 3.0 Tesla MRI scanner with a head coil. Echo-planar images (EPI) were acquired using a T2*-weighted gradient with whole-brain coverage (TR = 10.0 seconds, TA = 2.0 seconds, TE = 30 ms, flip angle = 90 degrees). Each volume consisted of thirty-two axial slices (slice thickness = 3.25 mm, inter-slice gap = 0.75 mm) acquired in ascending order (resolution 2.19 x 2.19 x 4.0mm, field of view = 280 x 224 x 128mm). Quadratic shim gradients were used to correct for magnetic field inhomogeneities within the anatomy of interest. Data were acquired in two consecutive scanning runs, each involving 96 whole-brain images. Run order was counter-balanced across participants. T1-weighted whole-brain structural images were also obtained in all subjects.

To avoid interference from scanner noise, functional data were acquired using a sparse sampling protocol (Hall et al., 1999). Stimuli were presented during 8-second intervals, interspersed with image acquisition periods (see Figure 6.1). Stimuli were presented using the E-prime software (Psychology Software Tools Inc., Pittsburgh, PA, USA). All participants wore cannulated earplugs that provided additional protection against scanner noise, while allowing transmission of the auditory stimuli.

The study consisted of six conditions. In the four auditory emotion conditions, the participants listened to non-verbal vocalisations of achievement/triumph, amusement, disgust, or fear. In the fifth condition, participants listened to the concatenated, spectrally rotated sounds, which provided an auditory baseline condition. In each auditory condition, the participants listened to three randomly selected stimuli from the given group of stimuli, presented during 2.5 second intervals.

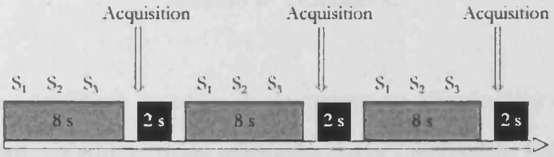

Fig. 6.1: Sparse scanning paradigm employed in the current study.

In any one of these conditions, the participant would only hear stimuli of one emotion, or the auditory baseline sounds. The instruction "Listen" was displayed on the video monitor during all of the auditory conditions. In the sixth condition, participants were cued to initiate a voluntary smiling movement by the appearance of the instruction "Smile" on the video monitor. Two very brief periods of relaxation were followed by resumption of the smiling movement, cued by the serial addition of two exclamation marks at the end of the "Smile" instruction. These cues appeared at 2.5 and 5 seconds after trial onset. Between each condition, a fixation cross was displayed on the video monitor. Consecutive trials were always from different conditions.

Imaging analysis

Data were pre-processed and analysed with Statistical Parametric Mapping software (SPM2, Wellcome Department of Imaging Neuroscience; http://www.fil.ion.ucl.ac.uk/spm/). Image pre-processing included realigning the EPI images to remove the effects of head movement, co-registration of T1 structural image to the mean EPI image, normalisation of EPI into Montreal Neurological Institute (MNI) standard stereotactic space using normalisation parameters derived from the co-registered T1-weighted image, and smoothing of the normalised EPI images using an 8 mm Gaussian filter.

The data were analysed using a random effects model. At the first level of analysis, individual design matrices were constructed for each participant, modelling the six experimental conditions. Movement parameters from the re-alignment step were included as additional regressors. Blood oxygen level dependent (BOLD) responses were modelled using a Finite Impulse Response (FIR) model of length 2.0 seconds, and order 1 (Gaab, Gabrieli & Glover, in press). Contrast images for each of the contrasts of interests were created at the first level for each participant, and entered into the second-level analyses.

Results & Discussion

A contrast comparing all auditory emotional conditions to the auditory baseline condition was computed in order to examine which areas were selectively involved in the processing of emotional sounds. Contrasts were also computed for the activation comparing each of the emotional auditory conditions with the auditory baseline, to examine which areas were involved in the perception of each of the different types of emotional sounds. This experiment also aimed to explore whether there are areas involved both in the perception of emotional sounds, and the emotional motor task condition. In order to study this issue, a contrast of all emotional sounds over baseline was again computed, this time using a mask consisting of areas involved in the motor production task. These analyses are discussed in turn below. All co-ordinates given are in Montreal Neurological Institute (MNI) standardised stereotaxic space, and all contrasts were computed using whole-brain correction with False Discovery Rate corrections at $p < 0.05$. Minimum cluster size was specified at 10 voxels. All Statistical Parametric maps (SPMs) are displayed on axial, sagittal, and coronal projections on an averaged image of the normalised brains of the 20 participants in the current experiment.

All auditory emotional conditions compared to the auditory baseline

In order to examine which areas were more active during the perception of emotional sounds compared to the auditory baseline, the contrast of all emotions over the auditory baseline was computed. Areas activated more for emotional sounds than the auditory baseline involved areas along the temporal lobes bilaterally, including extensive areas in the left and right STS (see Table 6.1). This is consistent with previous studies that have found bilateral STS activation during perception of non-verbal emotional sounds (Grandjean et al., 2005; Meyer, 2005). These peaks are also broadly consistent with the regions thought to be "voice selective" (Belin et al., 2000). The right pre-SMA was also found to be activated during the emotional sounds compared to the auditory baseline. Two previous studies using emotional facial expressions have found activation in the SMA bilaterally (Carr et al., 2003; Leslie et al., 2004), suggesting that this region may be involved in the mapping of affective perceptual input to appropriate motor patterns. Activation in the left insula was also increased, consistent with previous studies using facial (Carr et al., 2003), verbal (Wildgruber et al., 2002), and non-verbal vocal stimuli (Sander & Scheich, 2001; Morris et al., 1999) of emotion. This region has been shown to be involved in motor planning of

Tab. 6.1: Brain regions showing significant activation to all emotional sounds over the auditory baseline. Note: R = Right hemisphere, L = Left Hemisphere.

x	y	z	Cluster size	T	Z-value	Area
-62	-8	-2	316	8.81	5.49	L Superior temporal sulcus
54	10	-14	394	7.94	5.21	R Superior temporal sulcus (/gyrus)
60	-24	0	142	5.84	4.37	R Superior temporal sulcus
64	-36	14	92	5.67	4.29	R Superior temporal sulcus
-60	-40	8	73	5.57	4.24	L Superior temporal sulcus
8	12	58	27	5.06	3.98	R pre-supplementary motor area
-40	16	-4	27	4.79	3.83	L Anterior Insula

speech (Dronkers, 1996); the current finding suggests that its involvement in motor planning may extend to vocal production outside of the speech domain. The pattern of activation is illustrated in Figure 6.2.

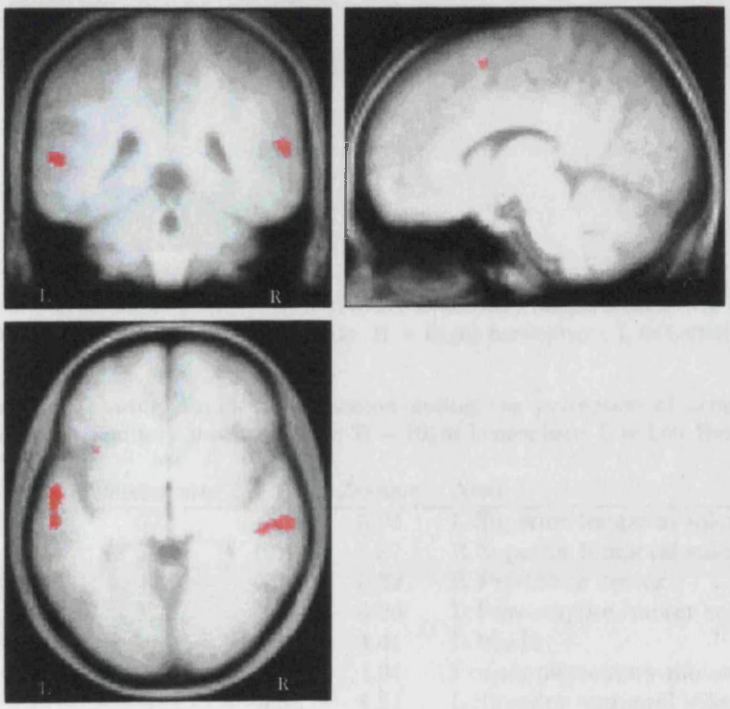

Fig. 6.2: An SPM of regions significantly activated to all emotions over auditory baseline, shown at x = 8.2, y = -39.7, z = -1.4. Note: R = Right hemisphere, L = Left Hemisphere.

Achievement/Triumph

In order to examine which areas of the brain were involved in the processing of achievement/triumph sounds, the contrast for these sounds over the auditory baseline was computed. Areas activated more for achievement/triumph sounds than the auditory baseline was similar to the pattern found for all emotions. Activated areas included temporal regions as well as areas involved in motor preparation (see Figure 6.3).

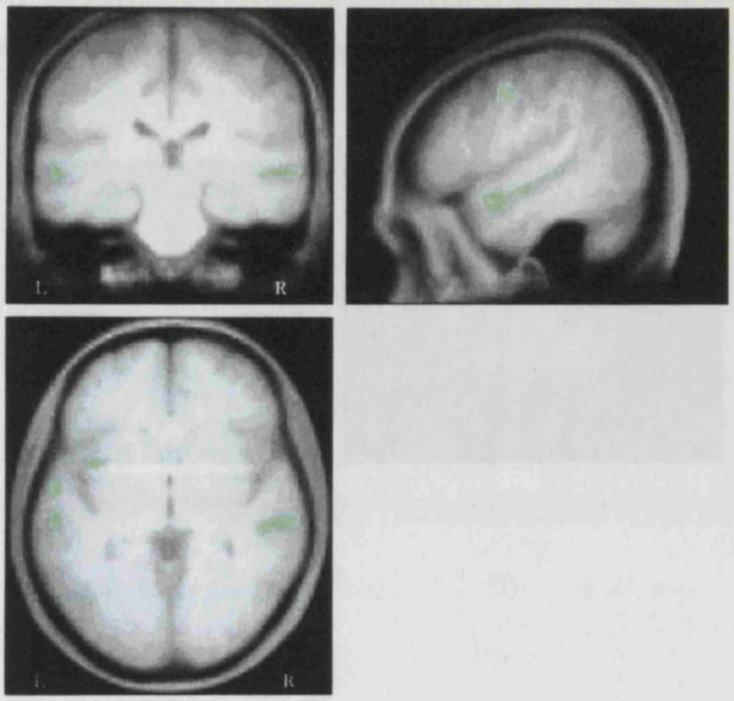

Fig. 6.3: An SPM of regions significantly activated to achievement/triumph sounds over auditory baseline, shown at x = 53.5, y = -25.6, z = 0. Note: R = Right hemisphere, L = Left Hemisphere.

Tab. 6.2: Brain regions showing significant activation during the perception of achievement/triumph sounds over the auditory baseline. Note: R = Right hemisphere, L = Left Hemisphere.

x	y	z	Cluster size	T	Z-value	Area
-62	-8	-2	602	10.36	5.93	L Superior temporal sulcus
62	-6	-10	680	10.12	5.87	R Superior temporal sulcus
54	2	44	84	7.97	5.22	R Pre-motor cortex
-48	-8	46	323	7.24	4.96	L Primary/pre- motor cortex
-42	6	4	116	5.94	4.41	L Insula
4	10	60	367	5.73	4.31	Pre-supplementary motor area
-56	-38	6	98	5.52	4.21	L Superior temporal sulcus
64	-38	14	20	5.19	4.05	R Superior temporal sulcus/gyrus

Specifically, the left and right STS, bilateral pre-motor cortex (bordering on the primary motor cortex in the left hemisphere), the pre-SMA, and the left insula were significantly more activated during the perception of achievement/triumph sounds than the auditory baseline (see Table 6.2). A noted above, these patterns of activation are largely consistent with previous work using emotional vocalisations in functional imaging.

Amusement

The contrast comparing activation during the perception of amusement sounds compared to the auditory baseline revealed activation in temporal regions in both hemispheres, located in the STS (see Figure 6.4).

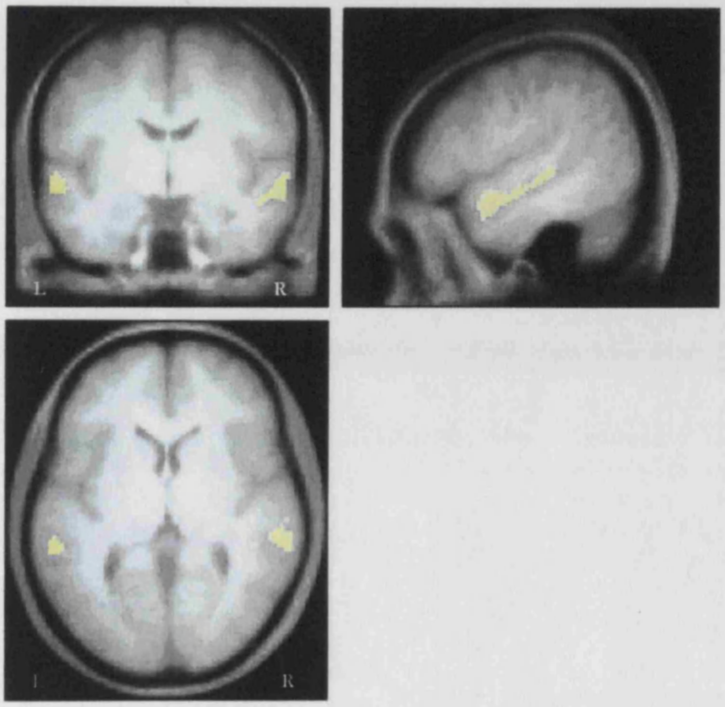

Fig. 6.4: An SPM of regions significantly activated during perception of amusement sounds over the auditory baseline, shown at x = 53.5, y = -6.2, z = 5. Note: R = Right hemisphere, L = Left Hemisphere.

Activation in the right posterior IFG was also increased (see Table 6.3). This area has been found to be activated during production of emotional facial expressions and has been called the human “mirror area” (Carr et al., 2003; Leslie et al., 2004).

Tab. 6.3: Brain regions showing significant activation to amusement sounds over the auditory baseline. Note: R = Right hemisphere, L = Left Hemisphere.

x	y	z	Cluster size	T	Z-value	Area
-56	2	-14	249	7.01	4.87	L Superior temporal sulcus
-60	-38	8	112	6.7	4.75	L Superior temporal sulcus
54	6	-14	731	6.4	4.62	R Superior temporal sulcus
46	12	22	110	5.56	4.23	R Posterior Inferior Frontal Gyrus (BA 44)

Disgust

Activation during the perception of disgust sounds compared to the auditory baseline was less extensive than for the other emotions. Activation was limited to the STS in both the left and right hemispheres (see Table 6.4).

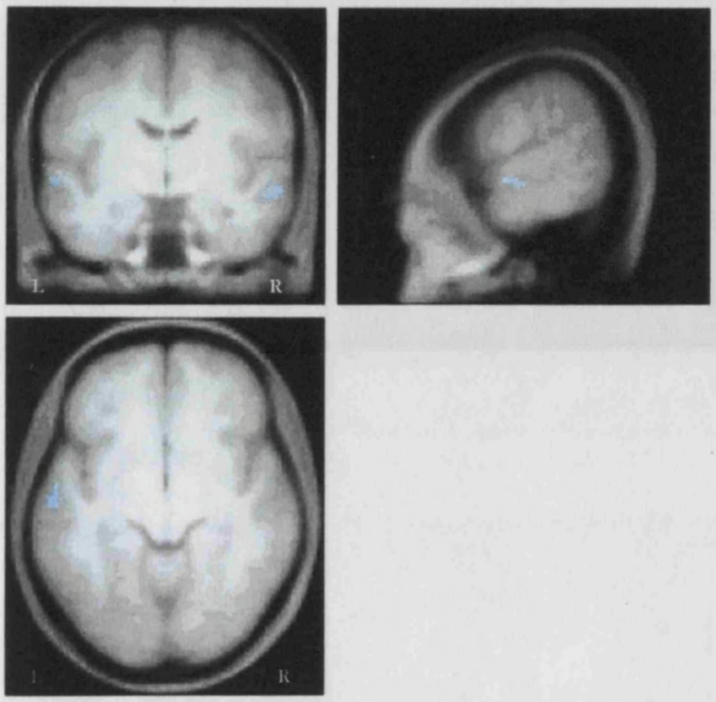

Fig. 6.5: An SPM of regions significantly activated during perception of disgust sounds over the auditory baseline, shown at x = -62.4, y = -7, z = -7. Note: R = Right hemisphere, L = Left Hemisphere.

The pattern of activation shown during the perception of disgust sounds over the auditory baseline is illustrated in Figure 6.5.

Tab. 6.4: Brain regions showing significant activation to disgust sounds over the auditory baseline. Note: R = Right hemisphere, L = Left Hemisphere.

x	y	z	Cluster size	T	Z-value	Area
60	-6	-10	21	6.86	4.81	R Superior temporal sulcus
-60	-8	-4	16	6.56	4.69	L Superior temporal sulcus

Fear

In order to examine which areas of the brain were involved in the processing of fear sounds, the contrast for these sounds over the auditory baseline was computed. Activation involved a network along the temporal lobes bilaterally, including STS and superior temporal gyrus (STG). The right pre-motor cortex was also activated, a region previous found to be activated during both viewing and imitation of emotional faces (Leslie et al., 2004). Pre-supplementary motor area, and right IFG (see Figure 6.6) were also activated during the perception of fear sounds, consistent with previous work suggesting that these regions are human "mirror areas" involved in mapping perceptual input of social signals to appropriate action maps (Carr et al., 2003; Leslie et al., 2004).

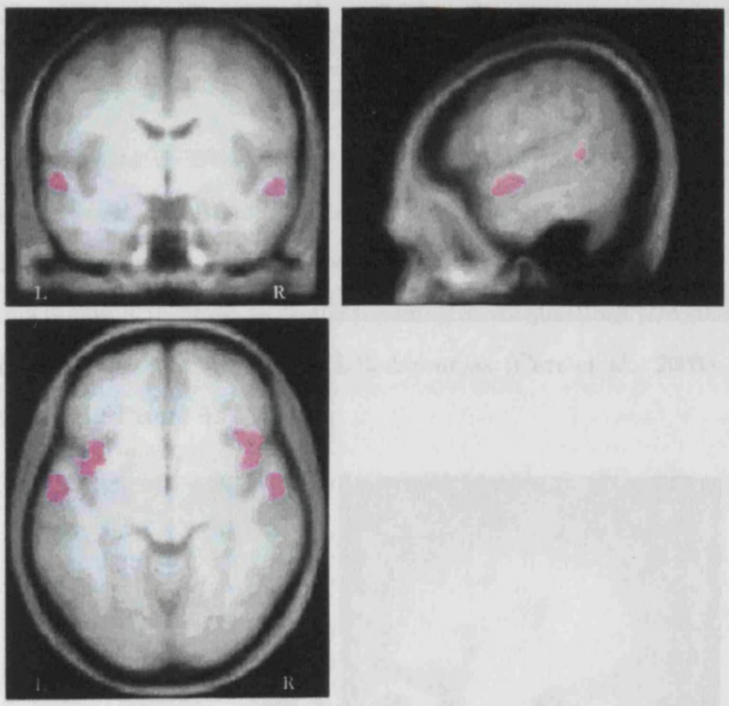

Fig. 6.6: An SPM of regions significantly activated during perception of fear sounds over the auditory baseline, shown at x = -57.9, y = -6.2, z = -5. Note: R = Right hemisphere, L = Left Hemisphere.

Activation was more extensive in the right hemisphere, with only the STS activated on the left. In the right hemisphere, the STS, the pre-SMA, the pre-motor cortex, and the inferior frontal gyrus were activated (see Table 6.5).

Common areas for listening to emotional sounds and an emotional motor task

To examine whether any common areas would be activated during the perception of emotional sounds and the performance of an emotional motor task, an additional analysis was included. The contrast of all emotions over the auditory baseline was computed, using the contrast of the motor condition over the auditory baseline as a mask. One of the areas activated by both listening to emotional sounds and the motor task was found in the right STS (see Table 6.6), close to the most

Tab. 6.5: Brain regions showing significant activation to fear sounds over the auditory baseline. Note: R = Right hemisphere, L = Left Hemisphere.

x	y	z	Cluster size	T	Z-value	Area
54	10	-14	870	7.76	5.15	R temporal pole
-58	-6	-4	472	6.77	4.77	L superior temporal gyrus
64	-24	-2	146	6.52	4.67	R medial temporal gyrus
-56	-44	10	87	5.67	4.29	L medial temporal gyrus
10	14	60	53	5.59	4.24	R Supplementary motor area
44	26	24	69	4.83	3.85	R Inferior frontal gyrus
58	-38	18	42	4.6	3.72	R superior temporal gyrus
42	8	48	10	4.36	3.59	R Pre-motor cortex
54	28	6	12	4.18	3.47	R Inferior frontal gyrus

anterior right peak specified by Belin et al. (2000) as "voice specific". Activation in the right pre-SMA was also increased, consistent with previous work with emotional facial stimuli (Leslie et al., 2004). Finally, the left insula was activated in both conditions. This is in line with accounts suggesting that this region is involved in motor planning of vocalisations (Dronkers, 1996), or may function as a link between motor regions and limbic areas (Carr et al., 2003). This pattern of activation is illustrated in Figure 6.7.

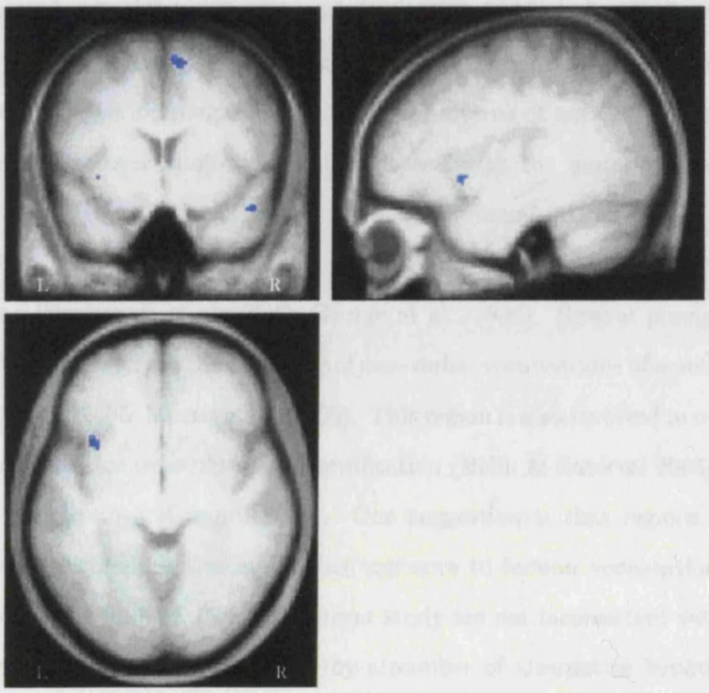

Fig. 6.7: An SPM of regions significantly activated during perception of all emotional sounds over the auditory baseline, masked with the contrast motor task over auditory baseline (also at FDR, $p < 0.05$), shown at x = -36.4, y = 12, z = -4. Note: R = Right hemisphere, L = Left Hemisphere.

Tab. 6.6: Brain regions showing significant activation during the perception of all sounds over the auditory baseline, masked with the contrast motor task over auditory baseline. R = Right hemisphere, L = Left Hemisphere.

x	y	z	Cluster size	T	Z-value	Area
54	6	-16	77	7.12	4.91	R Superior temporal sulcus
8	12	58	27	5.06	3.98	R Pre-supplementary motor area
-40	16	-4	27	4.79	3.83	L Anterior Insula

Conclusions

This study investigated neural responses to non-verbal emotional vocalisations. Areas found to be involved in the processing of emotional sounds included bilateral temporal regions, preparatory motor areas, the left anterior insula, and the right IFG. Each of these regions is now discussed in turn.

The Superior Temporal Sulcus

All of the contrasts, that is, all emotions together over the auditory baseline, with and without a mask of the motor task activation, as well as each of the individual emotions over the auditory baseline, activated the superior temporal sulcus. The patterns of activation were bilateral, and in one case (all emotions over auditory baseline masked with the motor task activation) right lateralised. This is consistent with previous studies on emotional prosody in speech, which have found bilateral (Grandjean et al., 2005; Kotz et al., 2003) or right-lateralised patterns of activation in temporal regions (Buchanan et al., 2000; George et al., 1996). Several previous studies have shown that the STS is involved in the processing of non-verbal vocalisations of emotions (Grandjean et al., 2005; Meyer et al., 2005; Morris et al., 1999). This region is also involved in other higher-level analyses of vocal stimuli, for example voice identification (Belin & Zatorre, 2003; von Kriegstein et al., 2003; von Kriegstein & Giraud, 2004). One suggestion is that regions of the STS are voice selective, thus selectively activating during exposure to human vocalisations (Belin et al., 2000; 2004). Although the findings from the current study are not inconsistent with this proposal, both these and other data could be explained by a number of alternative hypotheses, including involvement of this area in the processing of expertise and multi-modal social signals.

The link between perception and action in the processing of emotional vocalisations

One aim of this study was to examine whether the link between perception and action which has previously been found for speech (Fadiga et al., 2002; Watkins et al., 2003; Wilson et al., 2004) and emotional facial expressions (Carr et al., 2003; Leslie et al., 2004) would also exist for non-verbal vocal expressions of emotions. Several regions that are known to be involved in motor planning were activated in the current study. Both achievement/triumph and fear sounds activated the right pre-motor cortex, although the areas activated during the achievement/triumph sounds were more posterior, towards primary motor cortex. This could indicate that the activation seen during the achievement sounds is located further along the motor output pathway compared with fear. The regions activated during achievement/triumph sounds have previously been implicated in speech perception and production (Wilson et al., 2004).

An increase in activation was also found in the pre-SMA, during perception of achievement/triumph sounds, fear sounds, and in the contrast of all emotions over auditory baseline. Activation in this region was also found in the analysis of common areas of activation for emotional sounds and the motor task. Pre-SMA corresponds to area F6 in non-human primates, which has been implicated in higher-order aspects of complex motor control (Rizzolatti & Luppino, 2001). The left pre-SMA has previously been shown to respond to the perception and production of positive emotional facial expressions (Krolak-Salmon et al., 2006). In the current study, activation was right-lateralised in all cases except for achievement/triumph sounds, where they were bilateral. Caution should be employed in the interpretation of lateralisation, particularly with regard to an area so close to the midline of the brain, but this finding does imply that the involvement of the pre-SMA in affective perception-action links is not limited to the left side.

The right IFG was activated during sounds of amusement and fear. This region is the putative human homologue of the non-human primate mirror neuron area F5 (Rizzolatti & Arbib, 1998). In monkeys, neurons in this region respond when performing an action and when hearing sounds related to that same action (Kohler et al., 2002). This region is thought to encode motor prototypes, that is, representations of potential actions congruent with a particular stimulus (Rizzolatti & Luppino, 2001). The activation of this region during passive listening to emotional sounds of amusement and fear suggests that vocal communications may activate motor representations encoded in the right IFG. The motor representations likely correspond to a repertoire of orofacial gestures potentially appropriate to the emotional content of the perceived vocal stimulus. This process of auditory-motor interaction may be supported by the dorsal auditory pathway, which likely involves this region (Scott & Johnsrude, 2003; Warren, Wise & Warren, 2005). The IFG is also involved in perception and imitation of emotional facial expressions (Carr et al., 2003). This region may play a crucial role in social communication: A recent study found that autistic children demonstrate reduced activation in posterior IFG during observation and imitation of emotional facial expressions, and that the extent of the reduction correlates with measures of social dysfunction (Dapretto et al., 2006).

In the current study, the left anterior insula was also found to be involved in the processing of emotional sounds. This region was activated during achievement/triumph sounds, and in the contrast comparing all sounds to the auditory baseline, as well as during the motor task. The insula is well recognised from both human and animal studies as an area for processing visceral sensory stimuli (Augustine, 1996), and negative affect (Calder, Lawrence, et al., 2001; Damasio et al., 2000; Phillips et al., 1997). The anterior insula in particular has been shown to be involved in higher level processing of negative affect, such as financial risk taking (Kuhnen, & Knutson, 2005) and social rejection (Eisenberger, Lieberman & Williams, 2003). However, this region has been implicated in social communication of both positive and negative affect (Morris et al., 1999; Sander & Scheich, 2001; Wildgruber et al., 2002), suggesting that its involvement is not limited to negative emotional processing. The anterior insula likely serves a more generic role in affective processing, possibly linking together the evaluative functions of the limbic system with the motor system to enable appropriate responses (Carr et al., 2003). This is consistent with evidence showing that the anterior insula is involved in the motor production of facial social signals (Carr et al., 2003), and that this region is preferentially involved in speech production (Ackermann & Riecker, 2004; Blank, Scott, Murphy, Warburton & Wise, 2002; Dronkers, 1996).

Localisation of fear and disgust signals

The current study did not find that vocalisations of individual emotions were localised in areas that have previously been shown to be preferentially involved in the processing of particular emotions. Specifically, the amygdala has been shown to be important for the processing of fear (Adolphs et al., 1994; Adolphs et al., 1999; Breiter et al., 1996; Calder, Young, Rowland, et al., 1996; Calder, Lawrence, et al., 2001; Morris et al., 1996; 1998; Scott et al., 1997; Whalen, Rauch, Etcoff, McInerney, Lee & Jenike, 1998). It is worth noting that the current study used a 3T scanner: this type of scanner is associated with signal loss in some areas including the amygdala. Most of the work showing this pattern of localisation has been carried out with facial expressions of emotions, whereas the data from studies using vocal expressions of emotions is more equivocal. For example, a study by Phillips et al. (1998) found that fear sounds caused increased activation in the amygdala, while the same stimulus set was shown to deactivate the right amygdala in a later study (Morris et al., 1999). Sander and Scheich (2001) demonstrated increased activation in the amygdala during fear sounds, but also during laughter sounds, and Wildgruber et al. (2005) found no differential activation in the amygdala for fearful as compared to neutral speech. As discussed previously, neuropsychological data are also ambiguous with regards to whether the amygdala is crucial for the ability to process auditory signals of fear. A study by Scott et al. (1997) found that a patient with a bilateral amygdala lesion was impaired in the region of fear signals from auditory stimuli, a finding later replicated by Sprengelmeyer et al. (1999). However, a study by Anderson and Phelps (1998) found intact recognition of auditory fear signals in a patient with bilateral amygdala damage. They suggested that the amygdala is essential for the recognition of fear from visual, but not auditory stimuli, and that interactions between the amygdala and striatal areas form the crucial aspect of emotion perception from auditory cues. Although caution must be employed in the interpretation of null results, the findings from the current study showed no involvement of neither the amygdala nor striatal areas during perception of fear sounds, and thus cannot be said to support Anderson and Phelps' hypothesis or a model in which the amygdala alone plays the central role for the processing of auditory signals of fear.

There is also evidence to suggest that the processing of disgust is linked to specific regions in the brain. The insula and striatum in particular have been associated with the processing of disgust signals (Calder, Keane, Manes, Antoun, & Young, 2000; Calder, Lawrence, et al., 2001; Gray, Young, Barker, Curtis, & Gibson, 1997; Phillips et al., 1997; 1998).A meta-analysis of 55 functional imaging studies of emotion by Phan, Wager, Taylor and Liberzon (2002) pointed to the basal ganglia as the main site of activation during exposure to disgusting stimuli, but several studies have also found the insula to be selectively involved in the processing of disgust signals (Phillips et al., 1997; Phillips et al., 1998). Some work has shown a double dissociation between the regions involved in the processing of fear and disgust (see Calder et al., 2001). Again, the data from studies using vocal expressions of disgust have yielded less convincing results than the work that has used facial stimuli. A study by Sprengelmeyer et al. (1996) examined 13 patients with Huntington's disease, a disease that causes cell death in brain regions including the basal ganglia. The patients' emotion recognition was examined using forced-choice tasks with facial expressions of emotion from the Ekman and Friesen series (1976), and nonsense-speech sentences. The patients were impaired in the recognition of several emotions in both modalities, but the most severe impairments from both facial and vocal signals were found for disgust signals. Yet several studies using emotional vocalisations in functional imaging paradigm have failed to find any involvement of the insula and striatum in the processing of disgusted vocalisations (e.g., Phillips et al., 1998; Wildgruber et al., 2005), a result also yielded in the current study. The clear involvement of these regions in the processing of facial expressions of disgust does not seem to be mirrored in the auditory domain. Phillips et al. (1998) hypothesise that "activation of the anterior insula specifically in response to *facial* expressions of disgust may ... reflect the role of this structure in perception of oral gestures, in particular those related to expulsion or spitting out of unpleasant substances" (p. 1813, emphasis added). However, the stimuli used in both Phillips et al's and the current study are in fact closely associated with physical expulsion (e.g., sounds of retching), and it may be that rather than perception of oral gestures, the crucial aspect is to what extent the stimulus initiates production of such gestures. Recent work has shown that the anterior insula is more activated during the imitation of facial expressions compared to passive perception, suggesting that it is involved in motor planning (Carr et al., 2003).

Positive emotions in the brain

Research in affective neuroscience "has been somewhat preoccupied with the bad over the good" (Berridge, 2003, p.106), with much more research investigating negative emotions compared to positive affect. As discussed above, there is now a wealth of data on the neural processing of fear and disgust signals. In comparison, relatively little is known about how positive feelings are processed in the brain, despite the fact that many functional imaging studies have included a "happiness" condition (Phan et al., 2002). There seems to be substantial overlap between regions involved in the processing of happiness and disgust: A study of self-generated emotions found that happiness was associated with activation in the right insula, an area commonly associated with disgust processing (Damasio et al., 2000). A meta-analysis by Phan et al. (2002), found that nearly 70% of happiness induction studies reported an increase of activation in the basal ganglia, one of the areas thought to be central to disgust processing. The basal ganglia, particularly the striatum, has been shown to be responsive to the values associated with actions (Samejima, Ueda, Doya & Kimura, 2005). This may region seems to be processing the values associated with potential response actions to disgust and happiness signals (Phan et al., 2002). It is not clear why signals of these emotions would elicit a stronger evaluative action response than other emotions, such as fear and anger. The current study did not find any activation in the basal ganglia for amusement or achievement/triumph sounds, although other regions involved in motor planning were involved in the processing of both types of positive sounds.

Most previous studies have not differentiated between different types of positive affect. The two most stimulus sets arguably most commonly used in studies of affective neuroscience, the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 1999), and the Ekman and Friesen Pictures of facial affect (1976), each only include one positive category (the IAPS also only includes one class of negative stimuli). However, one signal associated with positive emotions, laughter, has received some attention from neuroscientists. As discussed previously, evidence from functional neuroimaging has shown an involvement of the amygdala and insula of the processing in laughter (Sander & Scheich, 2001). In addition, two studies using depth-electrodes have found that the left pre-SMA is involved in both perception and production of laughter (Fried et al., 1998; Krolak-Salmon et al., 2006). The current study found that the right IFG is involved in the processing in laughter sounds. Given that this region, as well as the insula and pre-SMA, is associated with motor responses to auditory input, this seems to imply that laughter is intrinsically contagious. This is consistent with the suggestion that laughter contagion is important for the establishment of social bonds between mother and child, and that it aids the child's development of a physical self-other distinction (Provine, 2000).

It is worth noting that laughter is likely a signal human that beings share with a number of other species. Primates laugh as a result of tickling and rough-and-tumble play, similar to that seen with human caregivers and children (Provine, 1996; 2000), and play- and tickle-induced vocalisations in rats are thought to have evolutionary relations to human laughter (Panksepp & Burgdorf, 1999). This ultrasonic laughter has also been observed in studies of rats' reproductive behaviour (Panksepp & Burgdorf, 2003), suggesting that it may be a more general signal of positive affect than laughter in humans.

In sum, this study demonstrates that passive perception of non-verbal emotional vocalisations robustly engages bilateral higher level auditory regions and a network of pre-motor cortical regions. This pattern suggests that hearing emotional sounds activates a preparatory motor response in the listener, similar to that seen during speech perception (Wilson et al., 2004) and facial expressions of emotions (Carr et al., 2003; Leslie et al., 2004). The perception of important social signals may be intrinsically bound to the production of appropriate responses, which may or may not constitute "mirroring" actions.

7. GENERAL CONCLUSIONS

This chapter summarises the findings produced in this thesis. It addresses the aims set out in Section 1.10, discussing the data, implications of this work, and outstanding questions for this area of research. It is concluded that both verbal and non-verbal vocalisations of positive and negative emotions are reliable communicative tokens with discrete expressions. Pitch and pitch variation play a central role in the perception of non-verbal emotional signals, whereas broad spectral cues are more important in the perception of emotional speech. A series of studies with a pre-literate Namibian tribe demonstrate that non-verbal vocalisations of emotions can be recognised across cultures, implying that they may be universal. Finally, the neural processing of non-verbal vocalisations of emotions involves a bilateral network including temporal areas and pre-motor regions, suggesting auditory-motor interactions in the perception of these sounds. Although more work is needed into vocal expressions of emotion in general and non-verbal vocalisations in particular, this thesis provides substantial evidence of their communicative value.

As outlined in Section 1.10, this thesis aimed to investigate non-verbal vocalisations of emotions. It examined the roles of categorical and dimensional factors, the contributions of different acoustic cues and the influence of culture. It also aimed to compare verbal and non-verbal vocalisations of emotions in terms of recognition and the contribution of acoustic cues, and to explore the neural processing of non-verbal vocalisations. Below, each of the questions set out Section 1.10 are addressed in turn, and future directions for this area of research are discussed.

Can vocalisations of positive emotions be reliably identified?

One of the aims of this thesis was to test the hypothesis that there is a set of positive emotions with distinct, recognisable expressions (Ekman, 1992b). The specific emotions investigated were achievement/triumph, amusement, contentment, sensual pleasure and relief (Ekman, personal communication). Employing a forced-choice task, Experiment 1a demonstrated that non-verbal vocalisations of these positive emotions could be reliably identified by naïve listeners. This finding was replicated in a different language group in Experiment 1b, and further in Experiment 2 in the context of negative emotional vocalisations. In these experiments, the correct label was most commonly selected by listeners for each of the stimulus types. These data demonstrate that non-verbal vocalisations of positive emotions can be reliably identified. Recognition was highest for amusement, achievement/triumph, and relief, and lower for sensual pleasure and contentment sounds. The errors made by the participants were also highly consistent across these studies, with the most common errors being the labelling of contentment sounds as sensual pleasure and sensual pleasure sounds as contentment. This pattern of results raises the possibility that contentment and sensual pleasure do not have entirely distinguishable vocal signals, and that they could comprise subsets of a broader emotion category, such as physical enjoyment. However, such interpretations must be made cautiously, given that the vocalisations of contentment and sensual pleasure were recognised at better-than-chance levels. In addition, the removal of contentment in Experiment 3 did not significantly improve recognition rates for sensual pleasure sounds, implying that these confusions may have arisen from variable stimulus quality rather than the relationship between sensual pleasure and contentment sounds.

Experiment 10 showed that these positive emotions could not only be identified from non-verbal sounds, but also from emotionally inflected speech. In this experiment, the broader emotion category "happiness" was included, in addition to the other, specific positive emotions. Although recognised at above-chance levels, recognition of happiness sounds was much poorer than for than all of the other emotions, positive and negative. This finding is consistent with a number of previous studies, which have found that expressions of happiness or joy are not well recognised from emotional speech (Banse & Scherer, 1996; Juslin & Laukka, 2001; Scherer et al., 1991). Experiment 10 demonstrated that expressions of specific positive emotions were substantially better recognised than expressions of happiness. Although recognition for expressions of all of the positive emotions was high for verbal stimuli, some were less well recognised than when non-verbal stimuli were used. Expressions of achievement/triumph, sensual pleasure, and relief were better recognised from non-verbal, as compared to verbal, stimuli, a reflection of the fact that overall recognition was also higher with non-verbal stimuli. In sum, the work of this thesis has demonstrated that expressions of positive emotions can be reliably identified from verbal and non-verbal vocalisations.

Are there vocal equivalents of the facial expressions of the basic emotions?

Several theorists have hypothesised that the basic emotions anger, disgust, fear, happiness, sadness, and surprise would have unique vocal, as well as facial, expressions (Ekman, 1992b; Tomkins, 1962). Previous studies have provided some support for this notion, but have either omitted some of the basic emotions (Laukka, 2004; Scherer, Banse & Wallbott, 2001), or have supplied only limited data (Scott et al., 1997). Thus, this thesis provides the first extensive investigation of the hypothesis of vocal correlates of the basic emotions, examining the ability of listeners to identify vocal expressions from both non-verbal and verbal vocalisations. Notably, this was carried out in the context of the extended set of positive emotions proposed by Ekman (1992b), which meant that the emotion happiness was not examined as one, but rather as several, emotions.

Experiment 2 showed that non-verbal vocal signals of all of the basic emotions could be reliably identified by naïve listeners. Recognition was highest for disgust sounds and lowest for expressions of surprise. Verbal signals of the basic emotions were also reliably recognised, as demonstrated in Experiment 10. However, as discussed above, recognition performance for speech expressing the broad category "happiness" was dramatically lower than stimuli from all of the other emotion categories. The individual positive emotions were, in contrast, recognised well. This pattern of results would seem to suggest that all of the negative and neutral basic emotions have distinct vocal expressions, but that happiness is more usefully fractionated into a set of distinct positive emotions, as suggested by Ekman (1992b).

Are non-verbal vocal expressions discrete or dimensional?

The question of whether emotions are best conceptualised as discrete or dimensional entities has dominated the debate in emotion research for decades. Although some previous studies have contrasted predictions from the two accounts using vocal stimuli (Bänziger & Scherer, 2005; Laukka, 2003), this thesis was the first to explore this issue in the context of non-verbal vocalisations of emotions. In Experiments 1a, 1b, and 2, participants rated the stimuli consistently, with each stimulus type being rated highest on its own scale. However, sounds of sensual pleasure were consistently rated as highly as, or higher than, contentment sounds on the contentment scale, suggesting that listeners had difficulty distinguishing between these two stimulus types. Contentment and sensual pleasure sounds were also commonly confused in the forced-choice tasks in Experiments 1a, 1b, and 2. Given that they were rated similarly in terms of arousal and pleasure this would seem to lend support for Russell's dimensional account. However, anger and fear, two other stimulus types that were rated very similarly for arousal and valence, were not commonly confused or rated highly on each other's scales. This suggests that participants' categorisation errors and ratings of the sounds did not consistently map onto the arousal and valence dimensions. Overall, these data lend support to the basic emotion rather than the dimensional account, as they imply that emotional signals are perceived as distinct, rather than fuzzy, entities, and that listeners' confusions and ratings do not consistently map onto their arousal and valence ratings.

The rating data were subjected to a principal components analysis to investigate whether, as predicted by dimensional accounts of emotion (Russell, 1980), the dimensions of arousal and valence underlie participants' perceptions of these sounds. The principal components analysis identified two factors underlying the listeners' ratings. These factors strongly correlated with listeners' ratings of valence and arousal, and accounted for a large amount of the variance in the rating data. This pattern would seem to suggest that these two dimensions help form listeners internal emotional space in the sense of how emotions relate to one another, even though listeners' categorisations errors and ratings did not consistently relate to those dimensions (see above). Notably, most of the variance in this analysis was accounted for by the factor which was correlated with participants' valence ratings. This is somewhat inconsistent with previous work suggesting that dimensional factors underlie vocal emotional signals, as these accounts have tended to emphasise the role of arousal (Bachorowski, 1999; Bänziger & Scherer, 2005), possibly because of their focus on a narrow set of acoustic cues.

In sum, these findings indicate that arousal and valence form part of listeners' internal emotional space, but do not consistently determine direct perceptual judgments, such as rating patterns and categorisation errors. This is in line with Laukka's (2004) suggestion that emotion dimensions may correspond to cognitive, but not perceptual, emotion constructs. This aspect of the dimensional accounts of emotions is not inconsistent with the basic emotion view of emotional signals as distinct, rather than fuzzy entities. The dimensions arousal and valence could underlie how emotional signals are perceived to relate to each other, in the sense of an internal emotional space. The same signals could nevertheless be perceived as communicating discrete emotional information, such as anger, fear or amusement.

Does the stimulus selection procedure affect recognition?

Most studies of emotional communication have used the best available stimuli (e.g., Banse & Scherer, 1996; Ekman et al., 1969; Scherer, Banse, & Wallbott, 2001; Schröder, 2003), while others have used stimulus sets matched for average recognition (Scott et al., 1997). Experiment 3 compared recognition accuracy for stimuli selected to match for accuracy and those selected for best recognition, based on recognition at pilot testing. This study demonstrated that, although recognition was significantly better than chance for both stimulus sets, recognition rates were significantly higher when the best available stimuli were used. Performance for all emotions improved, although the difference was greatest for the stimulus types that participants had most difficulty recognising when matched-level stimuli were used.

Is agreement inflated by the use of forced choice?

The majority of the tasks in this thesis — and arguably in research into emotional communication in general — have employed the forced choice paradigm. This methodology has been criticised for artificially inflating agreement, as participants are required to select one of the response alternatives offered (Russell, 1994). In response to this criticism, Frank and Stennett (2001) suggested offering "none of the above" as an additional response alternative, ensuring that participants are not forced to choose any of the emotional labels. In their study using emotional faces, they found that recognition rates were not significantly affected by this manipulation, which they interpreted as showing that the forced-choice paradigm provides an accurate reflection of participants' recognition. Experiment 4 in this thesis investigated whether the addition "none of the above" would affect listeners' recognition accuracy of non-verbal emotional vocalisations. This study showed that this manipulation did not affect participants' overall recognition of the sounds. When examining recognition rates for individual emotions, the addition of the "none" response option significantly reduced recognition rates only for expressions of relief and sadness. In line with Frank and Stennett's study, these data lend support to research using the forced-choice paradigm, showing that the recognition rates obtained with this methodology are not artificially inflated.

What is the role of different acoustic cues in the recognition of non-verbal expressions of emotions?

Several studies have investigated the acoustic bases of emotions in speech or pseudo-speech (Banse & Scherer, 1996, Bänziger & Scherer, 2005; Laukka, 2004; Murray & Arnott, 1993). Experiment 5 of this thesis attempted to investigate the roles of different acoustic cues in emotion perception from *non-verbal* vocalisations. The stimuli were acoustically manipulated using spectral rotations and noise-vocoding, to examine the contribution of different acoustic cues to emotion recognition.

A comparison of recognition of emotions from six-channel noise-vocoded and spectrally rotated stimuli showed that recognition was significantly better from spectrally rotated sounds than six-channel noise-vocoded sounds. This result demonstrates that pitch and pitch variation play an important role in the recognition of emotion from non-verbal vocalisations, as these cues were preserved in the spectrally rotated, but not the noise-vocoded, stimuli. This is consistent with research into emotion recognition from speech stimuli, which has emphasised the role of pitch cues (e.g., Bänziger & Scherer, 2005; Murray & Arnott, 1993), suggesting that identifying emotions from verbal and non-verbal vocalisations may rely to some extent on the use of similar acoustic cues.

Another type of acoustic cue that is thought to contribute to emotional communication in the voice is fine spectral detail. This cue captures properties of voice quality, such as tension, which are thought to contribute to the emotional quality of vocalisations (Murray & Arnott, 1993). Fine spectral detail was not present in either noise vocoded condition, nor in the spectrally rotated sounds. Recognition was impaired for all those conditions as compared to the original sounds, suggesting that voice quality is important for the recognition of emotion from non-verbal vocalisations.

The role of broad spectral cues was also investigated, by comparing six- and one-channel noise-vocoded stimuli. Speech intelligibility relies heavily on broad spectral variation (i.e., formants). One-channel noise-vocoded speech, where the broad spectral information is destroyed, is unintelligible (Shannon et al., 1995), whereas six-channel noise vocoded speech, where the broad spectral cues are preserved, is intelligible (Scott et al., 2000). If the same acoustic cues were important for identifying emotions as for understanding speech, emotions would have been better recognised from six- than one-channel noise-vocoded stimuli. However, there was no difference in performance between these two conditions. Although null results must always be interpreted with caution, this finding does suggest that the acoustic cues central to speech intelligibility do not contribute substantially to the identification of emotion in non-verbal vocalisations. In sum, the results from Experiment 5 showed that identification of emotion in non-verbal vocalisations relies mainly on pitch cues and fine spectral detail but not on formants, that are crucial for the intelligibility of speech.

What is the role of different acoustic cues in the recognition of emotional speech and how do they relate to the cues used to recognise emotions from non-verbal expressions of emotions?

Recent years have seen an increase in interest in emotional communication in the voice, almost exclusively focused on emotionally inflected speech or pseudo-speech. Several studies have investigated the acoustic basis of the communication of emotions in such utterances (e.g., Banse & Scherer, 1996, Bänziger & Scherer, 2005; Laukka, 2004; Murray & Arnott, 1993). As mentioned previously, research to date has had limited success in establishing specific patterns of acoustic cues that communicate discrete emotional states (Scherer, 1986; Juslin & Scherer, 2005), mainly due to the difficulty in measuring many acoustic cues (Scherer, 1986), and a (related) tendency to report only a small set of features, primarily relating to pitch (Juslin & Laukka, 2003). To avoid this problem, some studies have used acoustically manipulated speech to study the role of different acoustic features in emotional communication in speech (Ladd et al., 1985; Lieberman & Michaels, 1962). This was the approach used in this thesis. Examining emotion recognition from noise-vocoded and spectrally rotated speech, Experiment 11 aimed to investigate how acoustic factors affect the perceived emotion in emotional speech. The use of the same acoustic manipulations that were used in Experiment 5 with non-verbal stimuli allowed for a comparison of the acoustic cues important for emotion identification in verbal and non-verbal vocal stimuli.

Using non-verbal stimuli, Experiment 5 found that emotion was better recognised from spectrally rotated sounds than from six-channel noise-vocoded stimuli. This was interpreted to mean that pitch and pitch variation play an important role in the identification of emotion in non-verbal vocalisation. In Experiment 11, there was no difference in recognition for spectrally rotated and six-channel noise-vocoded stimuli, implying that recognition of emotion in speech does not rely heavily on pitch and pitch variation cues. This is somewhat surprising, as most previous studies have tended to emphasise the role of pitch cues in emotional speech (Banse & Scherer, 1996; Bänziger & Scherer, 2005; Murray & Arnott, 1993). This may be because previous work has utilised acoustic analysis rather than acoustic manipulations, especially given that global measures of pitch are relatively easy to measure. Studies that have used acoustic manipulations have tended to emphasise the complex interplay of several acoustic cues in emotional speech, of which pitch cues are only a part (Ladd et al., 1985; Lieberman & Michaels, 1962).

A feature known to play an important role in the comprehension of speech is broad spectral structure (Faulkner et al., 2001; Shannon et al., 1995). Although broad spectral structure was not found to be central to the perception of emotion in non-verbal stimuli, it did contribute to emotion recognition in speech. This was shown by six-channel noise-vocoded stimuli being better recognised than one-channel noise-vocoded stimuli for speech, but not non-speech stimuli. Similarly to speech comprehension, the recognition of emotion in speech thus relies to some degree on broad spectral structure. This may be due to an interaction with intelligibility processing occurring in the perception of emotional speech, which would not be present in the perception of non-verbal vocalisations. However, these results must be interpreted with some caution, given that one- and six-channel noise-vocoded stimuli differ in pitch information as well as broad spectral structure.

Previous work has also emphasised the role of voice quality, carried by fine spectral cues (Ladd et al., 1985; Scherer, 1986). Experiments 5 and 11 of this thesis showed that fine spectral cues play an important role in the perception of emotion in both verbal and non-verbal vocalisations. All of the acoustically manipulated stimuli were less well recognised than the original sounds, and this difference can likely be attributed in part to the absence of accurate fine spectral detail in all of the manipulated conditions. These cues are difficult to measure and have therefore received little attention in studies of emotional vocalisations; Acoustic manipulations may provide a fruitful way of investigating the role of voice quality in emotional signals in the voice.

Experiment 11 also replicated the finding from Experiment 10 that emotions are more easily recognised from non-verbal than the verbal stimuli. This was reflected in all of the acoustic conditions, with participants consistently performing better with non-verbal stimuli, although the extent of this advantage varied somewhat across acoustic conditions. In sum, broad spectral cues are important for emotion identification in speech, whereas pitch cues play a central role in the perception of emotion from non-verbal sounds. The decrease in performance seen in all manipulated conditions as compared to the original sounds suggests that fine spectral detail plays an important role for recognition of emotions in both verbal and non-verbal vocalisations.

Can acoustic analysis provide sufficient detail to statistically discriminate sounds from different emotional categories?

Statistical methods can successfully discriminate different emotions on the basis of basic perceptual cues, from both facial expressions and emotions in speech (Banse & Scherer, 1996; Calder et al., 2001; Laukka, 2004). Study 6 of this thesis investigated whether emotional expressions in non-verbal vocalisations could be discriminated on the basis of perceptual cues. Measurements of pitch cues, spectral cues and envelope information were used in a discriminant analysis in order to examine whether the information provided sufficient detail to discriminate between the different emotion categories. These analyses accurate at categorising the emotional vocalisations, demonstrating that the acoustic measurements provide sufficient information to successfully discriminate between stimuli from different emotional categories. An evaluation of the statistical models' performance showed that they was significantly better than chance not only in terms of overall performance but also in classifying stimuli from each of the emotional categories. The errors made by the models also mirrored those made by human listeners. These results demonstrate that non-verbal vocalisations can be accurately categorised from basic perceptual cues.

Can the acoustic features of the emotional sounds predict participants' perception of the sounds, as measured by emotional ratings?

Research has shown that listeners' perception of speech and nonsense-speech can be predicted by constellations of acoustic factors (Banse & Scherer, 1996: Laukka, 2004). One of the aims of this thesis was to map out what constellations of acoustic cues would be associated with listeners' perception of non-verbal vocalisations of emotion. Acoustic measurements of the sounds were used to predict participants' ratings on a set of emotional and dimensional scales. The analysis confirmed that participants' ratings on each of the emotion scales could be predicted from the acoustic measurements of the sounds, with a particular constellation of acoustic cues for each emotional scale (with the exception of contentment and sensual pleasure which were predicted by the same constellation). The significant constellation for each emotion scale consisted of cues including at least two of the three types of cues included in the acoustic analysis: amplitude envelope, pitch, and spectral cues, suggesting that the relationship between acoustic cues and perceived emotion is complex. This was also the case for listeners' ratings for arousal, in line with claims of an acoustic arousal dimension in vocal signals (Bachorowski, 1999; Bänziger & Scherer, 2005). The listeners' valence ratings could not be predicted by any constellation of acoustic cues, suggesting that there is not a strong relationship between the acoustic cues measured in this analysis and the perceived valence of non-verbal sounds. This finding is consistent with a study with emotional speech stimuli, where Laukka (2004) found that acoustic cues predicted markedly less variance for the participants' valence ratings compared to all of the other rating scales. In sum, the results of this study showed that listeners' ratings of emotions and arousal of non-verbal vocalisations can be predicted on the basis of acoustic cues, whereas valence cannot.

Can emotions vocalisations be communicated across cultures?

Only a small number of studies to date have investigated emotional communication in the voice cross-culturally (Juslin & Scherer, 2005). The consistent finding from these studies has been that vocal signals of emotions can be identified at levels that exceed chance across cultures (Elfenbein and Ambady, 2002b). However, these studies have largely ignored non-Western cultures. One aim of this thesis was to investigate to what extent emotional vocalisations are cross-culturally recognisable. Recognition and expression of non-verbal vocalisations was studied in a non-Western pre-literate culture. Three experiments demonstrated that non-verbal vocalisations of emotions can be recognised cross-culturally.

Experiment 7 showed that participants from a non-Western, pre-literate culture were able to reliably match non-verbal emotional vocalisations to brief emotion scenarios. Participants' performance was above chance overall, and for most of the emotions individually. Experiment 8 replicated this finding using a same-different task with participants from the same culture. Expressions of surprise were not reliably recognised in either Experiment 7 or 8, indicating that this group of participants did not recognise emotional vocalisations of surprise produced by Western speakers. This finding is consistent with previous cross-cultural work of emotional communication using facial expressions, which has found that members of non-literate cultures are unable to distinguish facial expressions of surprise from expressions of fear (Ekman and Friesen, 1971). This could indicate that surprise may be not a universally recognisable emotion when communicated using the face or voice.

In contrast to this, Experiment 9 demonstrated that members of the same non-Western group were able to *produce* non-verbal vocalisations of surprise, as well as all of the other emotions, which Western participants could reliably identify. Performance was better than chance both overall and for each emotion individually in this experiment, although recognition rates were lower than for stimuli produced by Western posers. This is in line with Elfenbein & Ambady's dialect account (2003), which holds that subtle differences in expressions between cultures mean that emotional expressions produced by members of a culture to which the receiver had not been exposed, are more difficult to recognise. In sum, the results from the investigation of cross-cultural recognition of emotions in this thesis indicate that non-verbal vocalisations of emotions are reliably recognisable even across vastly different cultures.

What is the role of spectral detail in speech intelligibility, emotion recognition, and speaker differentiation, and what are the relationships between these processes?

The role of spectral cues in speech perception is well established (Faulkner et al., 2001; Shannon et al., 1995; 2004). The role of these cues in the perception of emotion and voice identity has received less attention (but see Warren et al., 2006). One of the aims of this thesis was to examine the role of spectral information in speech perception, emotion, and identity differentiation, using acoustic degradation of spectral information. By gradually varying the degradation of the signal, the importance of this acoustic cue for the different processes can be established, in order to examine the extent to which these judgments rely on the same acoustic information. The results showed that listeners' performance improved with added channels for all three tasks: The improvement was most dramatic in the speech intelligibility task, and least pronounced in the emotion identification task. This suggests that broad spectral cues are crucial for speech intelligibility, but not sufficient for accurate recognition of emotions in speech, which may rely to a greater degree on voice quality cues. Speaker differentiation relied to an intermediate extent on spectral cues, and with little spectral resolution, listeners tend to judge two stimuli as perceptually similar. Although the difference in improvement was only statistically significant between the intelligibility and emotion tasks, there were strong trends towards a difference in the comparisons between the speaker differentiation and the other two tasks, and there was also no correlation between the improvements seen for individual participants for the different tasks in the current experiment These data indicate that the three tasks are showing three distinct relationships between performance and the amount of spectral information in the signal, lending some support to the hypothesis of independence of the processing of these different types of information in the voice (Belin et al., 2004).

What are the neural correlates of the perception of non-verbal vocalisations of emotions?

Experiment 13 examined how vocal expressions of emotion are processed in the brain, using an fMRI paradigm. This experiment showed that non-verbal emotional vocalisations, compared to an acoustically complex baseline, activate the STS bilaterally. This is in line with previous studies on emotional prosody in speech, which have found bilateral activation in temporal regions for speech or speech-like emotional vocalisations (Grandjean et al., 2005; Kotz et al., 2003) as well as non-verbal vocalisations of emotions (Meyer et al., 2005; Morris et al., 1999). It has been suggested that regions of the STS are "voice selective" (Belin et al., 2000; 2004). The findings from Experiment 13 are consistent with such a proposal, although it may be that this region is involved in higher-level analyses of social stimuli more generally. Listening to non-verbal vocalisations of emotions also activated a number of regions involved in motor preparation (see below).

Does the perception of non-verbal expressions of emotion activate areas involved in motor planning?

One aim of this thesis was to investigate links between auditory perception and action preparation in the context of emotional vocalisations. Previous studies of emotional facial expressions have shown that passive perception of emotional expressions can engage areas involved in motor planning (Carr et al., 2003; Leslie et al., 2004) and that speech perception can also elicit activation in areas associated with motor functions (Fadiga et al., 2002; Wilson et al., 2004). These findings suggest a close neural link between the perception of social signals and the preparation for responsive actions. Experiment 13 aimed to examine whether the link between perception and action previously found during passive perception of speech and facial expressions of emotions, would also exist for non-verbal vocal signals.

Several regions that are known to be involved in motor planning were activated during the passive perception of non-verbal vocalisations of emotions in Experiment 13. Both achievement/triumph and fear sounds activated the right pre-motor cortex and the pre-SMA. The right pre-motor cortex has previously been implicated in speech perception and production (Wilson et al., 2004), while the pre-SMA is involved in higher-order aspects of complex motor control (Rizzolatti & Luppino, 2001) and both the perception and production of positive emotional facial expressions (Krolak-Salmon et al., 2006). The pre-SMA was also activated when all of the emotions together were contrasted with the auditory baseline, and when this contrast was masked with the activation found during the motor task. These contrasts, as well as achievement/triumph sounds, also yielded activation in the left anterior insula. This region is involved in processing emotional signals (Morris et al., 1999; Sander & Scheich, 2001; Wildgruber et al., 2002), in particular negative affect (Calder, Lawrence, et al., 2001; Damasio et al., 2000; Phillips et al., 1997). It has been suggested that the anterior insula may serve as a link between the evaluative functions of the limbic system with the motor system to enable appropriate responses, consistent with its involvement in the motor production of facial social signals (Carr et al., 2003). Sounds of amusement and fear activated the right IFG, the putative human homologue of the non-human primate mirror neuron area F5 (Rizzolatti & Arbib, 1998). This region encodes representations of actions congruent with a particular stimulus (Rizzolatti & Luppino, 2001), and is involved in the perception and imitation of emotional facial expressions (Carr et al., 2003). In sum, passive perception of non-verbal emotional vocalisations robustly engages bilateral higher-level auditory regions as well as an extensive network of pre-motor cortical regions, likely eliciting a preparatory motor response in the listener. The perception of important social signals thus seems intrinsically bound to the production of appropriate responses.

Summary and future directions

This thesis is an investigation of vocal expressions of emotions, mainly focusing on non-verbal sounds. I hope to have demonstrated that such signals can be reliably identified by naïve listeners across cultural and linguistic boundaries, are acoustically distinct, and likely belong to discrete categories. This is the case for the "established" basic emotions (Ekman et al., 1969), as well as for a set of hypothesised positive basic emotions, including achievement/triumph, amusement, sensual pleasure, relief, and possibly contentment. The cues used to identify emotions in non-verbal vocalisations differ from the cues used when comprehending speech, and pitch and pitch variation play a central role. An additional set of studies using stimuli consisting of emotional speech demonstrated that these sounds can also be reliably identified, and that this is more dependent on broad spectral cues than emotion recognition from non-verbal stimuli. Finally, an fMRI study found that passive listening to non-verbal vocalisations of emotions activates a neural system of preparatory motor actions.

The research presented in this thesis addressed a number of questions set out in Section 1.10, but many of the issues explored in this thesis deserve to be examined more thoroughly in future studies. For example, how many of the set of 16 positive basic emotions that have more recently been proposed (Ekman, 2003) have recognisable vocal signals? What is the relationship between contentment and sensual pleasure? Is there any type of signal communicating surprise that is cross-culturally recognisable? And the role of different acoustic factors in the recognition of verbal and non-verbal signals is a topic that will benefit from the rapidly developing technology of acoustic manipulation and analysis. There are also a number of areas that this thesis did not touch on that would deserve attention. What is the phylogenetic continuity of non-verbal vocalisations? To what extent is the ability to recognise emotions from non-verbal vocalisations innate? What is the relationship between facial and vocal signals of emotion? As with much research, the studies in this thesis have perhaps raised more questions than they have answered. Nevertheless, I hope that this work has provided a first step towards a systematic, empirical study of non-verbal vocalisations of emotions.

References

Ackermann, H., & Riecker, A. (2004). The contribution of the insula to motor aspects of speech production: a review and a hypothesis. *Brain and Language, 89*, 320–328.

Adolphs, R., Damasio, H., & Tranel, D. (2002). Neural systems for recognition of emotional prosody: A 3-D lesion study. *Emotion, 2*, 23–51.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. R. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature, 372*, 669–672.

Adolphs, R., Tranel, D., Hamann, S., Young, A. W., Calder, A. J., Phelps, E. A., et al. (1999). Recognition of facial emotion in nine individuals with bilateral amygdala damage. *Neuropsychologia, 37*, 1111–1117.

Albas, D. C., McClusky, 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*, 481–490.

Anderson, A. K., & Phelps, E. A. (1998). Intact recognition of vocal expressions of fear following bilateral lesions of the human amygdala. *NeuroReport, 9*, 3607–3613.

Atkinson, A. P., Dittrich, W. H., Gemmell, A. J., & Young, A. W. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception, 33*, 717–746.

Augustine, J. R. (1996). Circuitry and functional aspects of the insular lobe in primates including humans. *Brain Research Review, 22*, 229–244.

Averill, J. R. (1980). A constructionist view of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research and experience: Vol. 1. Theories of emotion.* (pp. 305–339). New York: Academic Press.

Avis, J., & Harris, P. L. (1991). Belief-desire reasoning among Baka children: Evidence for a universal conception of mind. *Child Development, 62*, 460–467.

Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science, 8*, 53–57.

Bachorowski, J. A., & Braaten, E. B. (1994). Emotional intensity: Measurement and theoretical implications. *Personality and Individual Differences, 17*, 191–199.

Bachorowski, J. A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science, 6*, 219–224.

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

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication, 46*, 252–267.

Barnes, C. L., & Pandya, D. N. (1992). Efferent cortical connections of multimodal cortex of the superior temporal sulcus in the Rhesus monkey. *The Journal of Comparative Neurology, 318*, 222–244.

Beier, E. G., & Zautra, A. J. (1972). Identification of vocal communication of emotions across cultures. *Journal of Consulting and Clinical Psychology,39*, 166.

Belin, P., Fecteau, S. & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences, 8*, 129–134.

Belin, P., Fecteau, S., Charest, I., Nicastro, N., Hauser, M., & Armony, J.L. (2006). Human perception of animal affective vocalizations. *Poster presented at Organization of Human Brain Mapping Annual Meeting*, Florence, Italy.

Belin, P., & Zatorre, R.J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *NeuroReport, 14*, 2105–2109.

Belin, P., Zatorre, R. J., & Ahad P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research, 13*, 17–26.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad P., & Pike B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*, 309–312.

Berridge, K. C. (2003). Pleasures of the brain. *Brain and Cognition, 52*, 106–128.

Biehl, M., Matsumoto, D. Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of nonverbal Behavior, 21*, 3–21.

Birdwhistell, R. L. (1970). *Kinesics and context.* Philadelphia: University of Pennsylvania Press.

Blank, S. C., Scott, S. K., Murphy, K., Warburton, E., & Wise, R. J. S. (2002). Speech production: Wernicke, Broca and beyond. *Brain, 125*, 1829–1838.

Blesser B. (1972). Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *Journal of Speech and Hearing Research, 15*, 5–41.

Blonder, L. X., Bowers, D., & Heilman, K. M. (1991). The role of the right hemisphere in emotional communication. *Brain, 114*, 1115–1127.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences, 17*, 97–110.

Boersma, P., & Weenink, D. (2005). Praat: doing phonetics by computer. [Computer program]. http://www.praat.org/

Boucher. J. D., & Carlson O. E. (1980). Recognition of facial expression in three cultures. *Journal of Cross-cultural Psychology, 11*, 263–280.

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., et al. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron, 17*, 875–887.

Bruce, V. & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–27.

Buchanan, T.W., Lutz, K., Mirzazade, S., Specht, K., Shah, N. J., Zilles, K., et al. (2000). Recognition of emotional prosody and verbal components of spoken language: an fMRI study. *Cognitive Brain Research, 9*, 227–238.

Bullock, M., & Russell, J. A. (1985). Further evidence in preschoolers' interpretation of facial expressions. *International Journal of Behavioral Development, 8*, 15–38.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal components analysis of facial expressions. *Vision Research, 41*, 1179–1208.

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience, 3*, 1077–1078.

Calder, A. J., Lawrence, A. D., & Young, A. W. (2001). The neuropsychology of fear and loathing. *Nature Reviews Neuroscience, 2*, 352–363.

Calder, A. J., Young, A. W., Etcoff, N. L., Perrett, D. I., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition, 3*, 81–117

Calder, A. J., Young, A. W., Rowland, D., Perrett, D. I., Hodges, J. R., & Etcoff, N. L. (1996). Facial emotion recognition after bilateral amygdala damage: Differentially severe impairment of fear. *Cognitive Neuropsychology, 13*, 699–745.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6*, 641–651.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex, 11*, 1110–1123.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*, 649–657.

Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *NeuroImage, 14*, 427–438.

Campanella, S., Quinet, P., Bruyer, R., Crommelinck M., & Guerit, J.-M. (2002). Categorical perception of happiness and fear facial expressions: An ERP study. *Journal of Cognitive Neuroscience, 14*, 210–227.

Carr, L., Iacoboni, M., Dubeau, M-C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in human: 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.

Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology, 70*, 205–218.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, J. D., McWhinney, B., Flatt, M., & Provost, J. (1993). A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers, 25*, 257–271.

Connellan, J., Baron-Cohen, S., Wheelwright, S., Batki A., & Ahluwalia, J. (2001). Sex differences in human neonatal social perception. *Infant Behavior and Development, 23*, 113–118.

Coupland, N. J., Singh, A. J., Sustrik, R. A., Ting, P., & Blair, R.-J. R. (2003). Effects of diazepam on facial emotion recognition. *Journal of Psychiatry and Neuroscience, 28*, 452–463.

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics, 26*, 145–171.

Damasio, A. R., Grabowski, T. R., Bechara, A., Damasio, H., Ponto, L. L. B., Parvizi, J., et al. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience, 3*, 1049–1056.

Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., et al. (2006). Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience, 9*, 28–30.

Darwin, C. (1872/1998). *The Expression of Emotion in Man and Animals.* London: Harper Collins Publishers.

Dashiell, J. F. (1927). A new method of measuring reactions to facial expression of emotion. *Psychological Bulletin, 24*, 174–175.

Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature, 398*, 203–204.

Davidson, R. J. (1992). Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition, 20*, 125–151.

Davidson, R. J. (1998). Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition and Emotion, 12*, 307–330.

Davidson RJ. (2000). The functional neuroanatomy of affective style. In D. Lane & L. Nadel (Eds.), *Cognitive Neuroscience of Emotion* (pp. 371–88). New York: Oxford University Press

Davidson, R. J., Abercombie, H., Nitschke, J. B., & Putnam, K. (1999). Regional brain function, emotion and disorders of emotion. *Current Opinion in Neurobiology, 9*, 228–234.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*, 222–241.

Davitz, J. R. (1964). Personality, perceptual, and cognitive correlates of emotional sensitivity. In J. R. Davitz (Ed.), *The communication of emotional meaning* (pp 57–68). New York: McGraw-Hill.

de Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience, 7*, 242–249.

de Gelder, B., & Vroomen, J. (1996). Categorical perception of emotional speech. *The Journal of the Acoustical Society of America, 100*, 2818.

Denham, S. A., Zoller, D., & Couchoud, E. A. (1994). Socialization of preschoolers' emotion understanding. *Developmental Psychology, 30*, 928–936.

Dittrich, W. H., Troscianko, T., Lea, S. E., & Morgan, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. *Perception, 25*, 727–38.

Dronkers, N. F. (1996). A new brain region for coordinating speech articulation. *Nature, 384*, 159–161.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does Rejection Hurt? An fMRI Study of Social Exclusion. *Science, 10*, 290–292.

Ekman, P. (1972). *Emotions in the Human Face.* New York: Pergamon Press.

Ekman, P. (1992a). Are There Basic Emotions? *Psychological Review, 99*, 550–553.

Ekman, P. (1992b). An Argument for Basic Emotions. *Cognition and Emotion, 6*, 169–200.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin, 115*, 268–287.

Ekman, P. (2003). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life.* London: Weidenfeld and Nicolson.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behaviour: Categories, origins, usage, and coding. *Semiotica, 1*, 49–98.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*, 124–129.

Ekman, P., & Friesen, W. V. (1976). Pictures of facial affect. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V. O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology, 53*, 712–717.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science, 221*, 1208–1210.

Ekman, P., & Rosenberg, E. L. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* New York: Oxford University Press.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science, 164*, 86–88.

Elfenbein, H. A., & Ambady, N. (2002a). Is there an in-group advantage in emotion? *Psychological Bulletin, 128*, 243–249.

Elfenbein, H. A., & Ambady, N. (2002b). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*, 203–235.

Elfenbein, H. A. & Ambady, N. (2003). When familiarity breeds accuracy: Cultural exposure and facial emotion recognition. *Journal of Personality and Social Psychology, 85*, 276–290.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition, 44*, 227–240.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience, 15*, 399–402.

Faulkner, A., Rosen, S., & Wilkinson, L. (2001). Effects of the number of channels and speech-to-noise ratio rate of connected discourse tracking through a simulated cochlear implant speech processor. *Ear and Hearing, 22*, 431–438.

Feldman Barrett, L. & Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science, 8*, 10–14.

Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L. (2003). Mirror neuron responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *European Journal of Neuroscience, 17*, 1703–1714.

Fiez, J. A., Raichle, M. E., Balota, D. A., Tallal, P., & Petersen, S. E. (1996). PET Activation of posterior temporal regions during auditory word presentation and verb generation. *Cerebral Cortex, 6*, 1–10.

Frank, M. G., & Stennett, J. (2001). The forced-choice paradigm and the perception of facial expressions of emotion. *Journal of Personality and Social Psychology, 80*, 75–85.

Fredrickson, B. L. (1998). What Good Are Positive Emotions? *Review of General Psychology, 2*, 300–319.

Fried, I., Wilson, C. L., MacDonald, K. A., & Behnke, E. J. (1998). Electric current stimulates laughter. *Nature, 391*, 650.

Fu, Q-J., Chinchilla, S., Nogaki, G., & Galvin, J. J. (2005). Voice gender identification by cochlear implant users: The role of spectral and temporal resolution. *Journal of the Acoustical Society of America, 118*, 1711–1718.

Gaab, N., Gabrieli, J., & Glover, G. (in press). Assessing the influence of scanner background noise on auditory processing I: An fMRI study comparing three experimental designs with varying degrees of scanner noise. *Human Brain Mapping.*

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119*, 593–609.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise in cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*, 191–197.

George, M. S., Parekh, P. I., Rosinsky, N., Ketter, T. A., Kimbrell, T. A., Heilman, K. M., Herscovitch, P., et al. (1996). Understanding emotional prosody activates right hemisphere regions. *Archives of Neurology, 53*, 665–670.

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K.R., et al. (2005). The voices of wrath: brain responses to angry prosody in meaningless speech. *Nature Neuroscience, 8*, 145–146.

Grandjean, D., & Scherer, K.R. (2006). An electroencephalographic investigation of the temporal unfolding of emotion-constituent appraisal. *Journal of Cognitive Neuroscience, Supplement 1*, 182.

Gray, J. M., Young, A. W., Barker, W. A., Curtis, A., & Gibson, D. (1997). Impaired recognition of disgust in Huntington's disease gene carriers. *Brain, 120*, 2029–2038.

Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories.* Chicago: University of Chicago Press.

Haidt, J. & Keltner, D. (1999). Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition. *Cognition and Emotion, 13*, 225–266.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85*, 845–857.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliot, M. R., et al. (1999). "Sparse" temporal sampling in auditory fMRI. *Human Brain Mapping, 7*, 213–223.

Hall, J. A., & Matsumoto, D. (2004). Gender differences in judgments of multiple emotions from facial expressions. *Emotion, 4*, 201–206.

Hornak, J., Bramham, J., Rolls, E. T., Morris, R. G., O'Doherty, J., Bullock, P. R., et al. (2003). Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain, 126*, 1691–1712.

Hornak, J., Rolls, E. T., & Wade, D. (1996). Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia, 34*, 247-261.

Iidaka, T., Omori, M., Murata, T., Kosaka, H., Yonekura, Y., Okada, T., et al. (2001). Neural interaction of the amygdala with the prefrontal and temporal cortices in the processing of facial expressions as revealed by fMRI. *Journal of Cognitive Neuroscience, 13*, 1035–1047.

Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., et al. (1997). Vocal identification of speaker and emotion activates differerent brain regions. *Neuroreport, 8*, 2809–2812.

Izard, C. (1971). *The Face of Emotion.* New York: Appleton-Century-Crofts.

Izard, C. E. (1977). *Human Emotions.* New York: Plenum Press.

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (2nd edition),(pp 220–235). New York: Guilford Press.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion, 1*, 381–412.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*, 770–814.

Juslin, P. N., & Scherer, K. R. (2005). Vocal expressions of affect. In J. A. Harrigan, R. Rosenthal & K. R. Scherer (Eds.). *The new handbook of methods in nonverbal behavior research* (pp. 65–136). New York: Oxford University Press.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*, 4302–4311.

Karama, S., Lecours, A. R., Leroux, J. M., Bourgouin, P., Beaudoin, G., Joubert, S., et al. (2002). Areas of brain activation in males and females during viewing of erotic film excerpts. *Human Brain Mapping, 16*, 1–13.

Karpf, A. (2006). The human voice. London: Bloomsbury.

Kohler, E., Keysers, C., Umit, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297*, 846–848.

Kotsoni, E., de Haan, M., & Johnson, M. H. (2001). Categorical perception of facial expressions by 7–month-old infants. *Perception, 30*, 1115–1125.

Kotz, S.A., Meyer, M., Alter, K., Besson, M., von Cramon, D. Y., & Friederici, A. D. (2003). On the lateralization of emotional prosody: An event-related functional imaging study. *Brain and Language, 86*, 366–376.

Kring, A. M., Barrett, L. F., & Gard, D. E. (2003). On the broad applicability of the affective circumplex: Representations of affective knowledge among schizophrenia patients. *Psychological Science, 14*, 207–214.

Krolak-Salmon, P., Hénaff, M-A., Vighetto, A., Bauchet, F., Bertrand, O., Mauguire, F., et al. (2006). Experiencing and detecting happiness in humans: The role of the supplementary motor area. *Annals of Neurology, 59*, 196–199.

Kucharska-Pietura, K., David, A. S., Masiak, M., & Phillips, M. L. (2005). Perception of facial and vocal affect by people with schizophrenia in early and late stages of illness. *British Journal of Psychiatry, 187*, 523–528.

Kuhnen, C. M., & Knutson, B. (2005). The neural basis of financial risk taking. *Neuron, 47*, 763–770.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *The Journal of the Acoustical Society of America, 78*, 435–444.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). International Affective Picture System (IAPS): Technical Manual and Affective Ratings. Gainsville, FL, The Centre for Research in Psychophysiology, University of Florida.

Laukka, P. (2003). Categorical perception of emotion in vocal expression. *Annals of the New York Academy of Sciences, 1000*, 283–287.

Laukka, P. (2004). Vocal expression of emotion. Discrete-emotions and Dimensional Accounts. Unpublished PhD thesis, University of Uppsala, Sweden.

Laukka, P. (2005). Categorical perception of vocal emotion expressions. *Emotion, 5*, 277–295.

Lazarus, R. S. (1991). *Emotion and adaptation.* New York: Oxford University Press.

Leslie, K. R., Johnson-Frey, S. H., & Grafton, S. T. (2004). Functional imaging of face and hand imitation: towards a motor theory of empathy. *NeuroImage, 21*, 601–607.

Levenson, R. W., Carstensen, L. L., Friesen, W. V., & Ekman, P. (1991). Emotion, physiology, and expression in old age. *Psychology and Aging, 6*, 28–35.

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial expression generates emotion-specific nervous system activity. *Psychophysiology, 27*, 363–384.

Levenson, R. W., Ekman, P., Heider, K., & Friesen, W. V. (1992). Emotion and autonomic nervous system activity in the Minangkabau of West Sumatra. *Journal of Personality and Social Psychology, 62*, 972–988.

Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *The Journal of the Acoustical Society of America, 34*, 922–927.

Macmillan, N. A., & Credman, C. D. (1996) *Detection theory: A users guide.* New York: Cambridge University Press.

Matsumoto, D. (1989). Cultural influences on the perception of emotion. *Journal of Cross-Cultural Psychology, 20*, 92–105.

Matsumoto, D. (2002). Methodological requirements to test a possible in-group advantage in judging emotions across cultures: Comment on Elfenbein and Ambady (2002) and evidence. *Psychological Bulletin, 128*, 236–242.

Matsumoto, D., & Assar, M. (1992). The effects of language on judgements of universal facial expressions of emotions. *Journal of Nonverbal Behavior, 16*, 85–99.

Matsumoto, D., & Ekman, P. (1988). Japanese and Caucasian facial expressions of emotion (JACFEE). [Slides]. Department of Psychology, San Francisco State University.

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience, 9*, 605–610.

Medin, D. L., & Atran, S. (1999). (Eds.) *Folkbiology.* Cambridge, MA: MIT Press.

Meyer, M., Zysset, S., von Cramon, D. Y., & Alter, K. (2005). Distinct fMRI responses to laughter, speech, and sounds along the human peri-sylvian cortex. *Cognitive Brain Research, 24*, 291–306.

Mitchell, R.L.C., Elliot, R., Barry, M., Cruttenden, A., & Woodruff, P. W. R. (2003). The neural response to emotional prosody, as revealed by functional magnetic resonance imaging. *Neuropsychologia, 41*, 1410–1421.

Morris, J. S., Frith, C. D., Perrett, D. I., Rowland, D., Young, A. W., Calder A. J., et al. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature, 383*, 812–815.

Morris, J. S., Friston, K. J., Bchel, C., Frith, C. D., Young, A. W., Calder, A. J., & Dolan, R. J. (1998). A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain, 121*, 47–57.

Morris, J. S., Scott, S. K., & Dolan, R. J. (1999). Saying it with feeling: Neural responses to emotional vocalizations. *Neuropsychologia, 37*, 1155–1163.

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*, 1097–1108.

Murray, I. R., Arnott, J. L., & Rohwer, E. A. (1996). Emotional stress in synthetic speech: Progress and future directions. *Speech Communication, 20*, 85–91.

Nakamura, M., Buck, R., & Kenny, D. A. (1990). Relative contributions of expressive behavior and contextual information to the judgment of the emotional state of another. *Journal of Personality and Social Psychology, 59*, 1032–1039.

Norenzayan, A., & Heine, S. J. (2005). Psychological Universals: What are they and how can we know? *Psychological Bulletin, 131*, 763–784.

Ortony, A., & Turner, T.J. (1990). What's basic about basic emotions? *Psychological Review, 97*, 315–331.

Panksepp, J., & Burgdorf, J. (1999). Laughing rats? Playful tickling arouses high frequency ultrasonic chirping in young rodents. In S. Hameroff, D. Chalmers, & A. Kazniak (Eds.), *Toward a science of consciousness III* (pp. 231–244). Cambridge, MA: MIT Press.

Panksepp, J., & Burgdorf, J. (2003). "Laughing" rats and the evolutionary antecedents of human joy? *Physiology and Behavior, 79*, 533–547.

Pell, M. D. (2006). Cerebral mechanisms for understanding emotional prosody in speech. *Brain and Language, 96*, 221–234.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage, 16*, 331–348.

Phelps, E.A., LaBar, K.S., Anderson, A.K., O'Connor, K.J., Fulbright, R.K., & Spencer, D. D. (1998). Specifying the contributions of the human amygdala to emotional memory: A case study. *Neurocase, 4*, 527–540

Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, S. C., et al. (1998). Neural responses to facial and vocal expressions of fear and disgust. *Proceedings of the Royal Society of London, 265*, 1809–1817.

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrews, C., Calder, A. J., et al. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature, 389*, 495–498.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis.* New York: Harper and Row.

Poeppel, D., Guillemin, A., Thompson, J., Fritz, J., Bavelier, D., & Braun, A. (2004). Auditory lexical decision, categorical perception and FM direction discrimination differentially engage left and right auditory cortex. *Neuropsychologia, 42*, 183–200.

Provine, R. R. (1996). Contagious yawning and laughter: Significance for sensory feature detection, motor pattern generation, imitation, and the evolution of social behaviour. In C.M.Heyes & B. G. Galef (Eds.), *Social Learning in Animals: The Roots of Culture* (pp. 179–208). New York: Academic Press.

Provine, R. R. (2000). *Laughter: A Scientific Investigation.* London: Faber and Faber.

Reisenzein, R. (2000). Exploring the strength of association between the components of emotion syndromes: The case of surprise. *Cognition and Emotion, 14*, 1–38.

Richardson, U., Thomson, J. M., Scott, S. K., & Goswami, U. (2004) Auditory processing skills and phonological representation in dyslexic children. *Dyslexia, 10*, 215–233.

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences, 21*, 188–194.

Rizzolatti, G., & Luppino, G. (2001). The Cortical Motor System. *Neuron, 31*, 889–901.

Roberson, D., Davidoff, J., Davies, I. R., & Shapiro, L. R. (2005). Color categories: evidence for the cultural relativity hypothesis. *Cognitive Psychology, 50*, 378–411.

Robin, O., Rousmans, S., Dittmar A., & Vernet-Maury, E. (2003). Gender influence on emotional responses to primary tastes. *Physiology & Behavior, 78*, 385–393.

Rodden, A., Wild, B., Erb, M., Titze, M., Ruch, W., & Grodd, W. (2001). Humour, laughter and exhilaration studied with functional Magnetic Resonance Imaging (fMRI). *NeuroImage, 13*, 466.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London, 336*, 367–373.

Rosenberg, E. L., & Ekman, P. (1994). Coherence between expressive and experiential systems in emotion. *Cognition and Emotion, 8*, 201–229.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178.

Russell, J. A. (1993). Forced-choice response format in the study of facial expression. *Motivation and Emotion, 17*, 41–51.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. *Psychological Bulletin*, 115, 102–141.

Russell, J. A., Bachorowski, J-A., & Fernández-Dols, J-M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology, 54*, 329–349.

Russell, J. A., & Bullock, M. (1986). On the dimensions preschoolers use to interpret facial expressions of emotion. *Developmental Psychology, 22*, 97–101

Russell, J. A., & Fehr, B. J. (1994). Fuzzy concepts in a fuzzy hierarchy: varieties of anger. *Journal of Personality and Social Psycholy, 67*, 186–205.

Russell, J. A. & Fehr, B. (1987), Relativity in the perception of emotion in facial expressions. *Journal of Experimental Psychology: General, 116*, 223–237.

Sabini, J., & Silver, M. (1997). In defense of shame: shame in the context of guilt and embarrassment. *Journal for the Theory of Social Behaviour, 27*, 1–15.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science, 310*, 1337–1340.

Sander, D., Grandjean, D., Pourtois, G., Schwartz, S., Seghier, M., Scherer, K., et al. (2005). Emotion and attention interactions in social cognition: brain regions involved in processing anger prosody. *NeuroImage, 28*, 848–858.

Sander, K., & Scheich, H. (2001). Auditory perception of laughing and crying activates human amygdala regardless of attentional state. *Cognitive Brain Research, 12*, 181–198.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin, 99*, 143–165.

Scherer, K. R. (1994). Affect bursts. In S. van Goozen, N. E. van de Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 161–196). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1989). Vocal correlates of emotion. In H. Wagner & A. Manstead (Eds.), *Handbook of Psychophysiology: Emotion and Social Behaviour*(pp 165–197). London: Wiley.

Scherer, K. R. (1997). The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology, 73*, 902–922.

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* (pp.92–120). New York: Oxford University Press.

Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication, 40*, 227–256.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross Cultural Psychology, 32*, 76–92.

Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (2001). Vocal cues in emotion encoding and decoding. *Motivation and Emotion, 15*, 123–148.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences, 10*, 24–30.

Schlosberg, H. (1941). A scale for the judgment of facial expression. *Journal of Experimental Psychology, 29*, 497–510.

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology, 44*, 229–237.

Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication, 40*, 99–116.

Scott, S. K., Blank, S. C., Rosen, S., & Wise, R. J. S. (2000) Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123*, 2400–2406.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences, 26*, 100–107.

Scott, S. K., Young, A. W., Calder, A. J., Hellawell, D. J., Aggleton, J. P., & Johnson, M. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature, 385*, 254–257.

Seltzer, B., & Pandya, D. N. (1978). Afferent connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. *Brain Research, 149*, 1–24.

Shannon R. V., Fu, Q-J., & Galvin, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngologica, 552*, 50–54.

Shannon, R. V., Zeng, F-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303–304.

Smith, Z. M., Delgutte, B., & Oxenham, A.J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416*, 87–90.

Smith, M., & Walden, T. (1998). Developmental trends in emotion understanding among a diverse sample of African-American preschool children. *Journal of Applied Developmental Psychology, 19*, 177–197.

Sprengelmeyer, R., Young, A. W., Calder, A. J., Karnat, A., Lange, H., Hornberg, V., et al. (1996). Perception of faces and emotions: loss of disgust in Huntington's disease. *Brain, 119*, 1647–1665.

Sprengelmeyer, R., Young, A. W., Shroeder, U., Grossenbacher, P. G., Federlein, J., Buttner, T., et al. (1999). Knowing no fear. *Proceedings of the Royal Society of London, 266*, 2451–2456.

Terracciano, A., Abdel-Khalek, A. M., Ádám, N., Adamovová, L., Ahn, C.-k., Ahn, H.-n., Alansari, B. M., et al. (2005). National character does not reflect mean personality trait levels in 49 cultures. *Science, 310*, 96–100.

Thompson, W. F., & Balkwill, L-L. (2006). Decoding speech prosody in five languages. *Semiotica, 158*, 407–424.

Toivanen, J., Vyrynen, E., & Seppnen, T. (2005). Gender differences in the ability to discriminate emotional content from speech. Proceedings FONETIK 2005, 119–122.

Tomkins, S. S. (1955). Consciousness and the unconscious in a model of the human being. *Proceedings of the 14th International Congress of Psychology* (pp. 160-161). Amsterdam; North-Holland.

Tomkins, S. S. (1962). *Affect, imagery, consciousness. (Vol. I).* New York: Springer.

Tomkins, S. S., & McCarter, R. (1964). What and where are the primary affects? Some evidence for a theory. *Perceptual and Motor Skills, 18*, 119–158.

Tyler, R. S., & Summerfield, A. Q. (1996). Cochlear implantation: relationships with research on auditory deprivation and acclimatization. *Ear and Hearing, 17*, 38-50.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perrett, D. I. (1997). Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition, 63*, 271–313.

van Bezooijen, R., Otto, S. A., & Heenan, T. A. (1983). Recognition of vocal expressions of emotion. *Journal of cross-cultural psychology, 14*, 387–406.

Vernet-Maury, E., Alaoui-Ismali, O., Dittmar, A., Delhomme. G., & Chanel, J. (1999). Basic emotions induced by odorants: a new approach based on autonomic pattern results. *Journal of the Autonomic Nervous System, 75*, 176–183.

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A-L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17*, 48–55.

von Kriegstein, K., & Giraud, A-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage, 22*, 948- 955.

Warren, J. D., Scott, S. K., Price, C. J., & Griffiths, T. D. (2006). Human brain mechanisms for the early analysis of voices. *NeuroImage, 31*, 1389–1397.

Warren, J. E., Wise, R. J. S., & Warren, J. D. (2005). Sounds do-able: auditory-motor transformations and the posterior temporal plane. *Trends Neurosciences, 28*, 636–643.

Watkins, K. E., Strafella, A.P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*, 989–994.

Watson, S. G. (1972). Judgement of emotion from facial and contextual cue combinations. *Journal of Personality and Social Psychology, 24*, 334–342.

Watson, D., & Tellegen, A. (1985). *Psychological Bulletin, 98*, 219–235.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998) Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience, 18*, 411–418.

Wildgruber, D., Pihan, H., Ackermann, H., Erb, M., & Grodd, W. (2002). Dynamic brain activation during processing of emotional intonation: influence of acoustic parameters, emotional valence, and sex. *NeuroImage 15*, 856–869.

Wildgruber, D., Hertrich, I., Riecker, A., Erb, M., Anders, S., Grodd, W., et al. (2004). Distinct frontal regions subserve evaluation of linguistic and emotional aspects of speech intonation. *Cerebral Cortex, 14*, 1384–1389.

Wildgruber, D., Riecker, A., Hertrich, I., Erb, M, Grodd, W., Ethofer, T., et al. (2005). Identification of emotional intonation evaluated by fMRI. *Neuroimage, 24*, 1233–1241.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*, 701–702.

Zatorre, R. J. (2001). Neural specialization for tonal processing. *Annals of the New York Academy of Sciences, 930*, 193–210.

APPENDIX

A. SCENARIOS FOR EMOTIONS AND DIMENSIONS.

Scenarios for positive and negative emotions, and the dimensions arousal and valence.

Emotion	Scenario
Achievement	You get a phone call offering you a job you really want
Amusement	You are being tickled and find it really funny
Anger	Someone is deliberately very rude to you
Arousal	Minimum: You are feeling sleepy Maximum: You are very awake and alert
Contentment	You are sitting on the beach watching the sunset
Disgust	You put your hand in vomit
Fear	Someone suddenly taps on your shoulder in a dark alleyway
Pleasure	Your boyfriend/girlfriend is touching you in a sensual way
Relief	You thought you had lost your keys but find them again
Sadness	You find out that someone close to you has died
Surprise	You find out you have been elected as an honorary citizen of a country you have never heard of
Valence	Positive: You are having an ecstatic experience Negative: You are experiencing trauma or extreme fear

B. CHI VALUES FOR RECOGNITION OF NON-VERBAL VOCALISATIONS FOR EACH EMOTION IN EXPERIMENT 2.

Chi values for recognition of positive and negative emotional vocalisations. All analyses significant at $p < 0.0001$, Bonferroni corrected for 10 comparisons. Degrees of freedom for all analyses = 9.

Emotion	Chi
Achievement/Triumph	1009.9
Amusement	1087.4
Anger	733.1
Contentment	422.3
Disgust	1550.8
Fear	664.5
Pleasure	716.4
Relief	1286.0
Sadness	794.1
Surprise	481.3

C. CHI VALUES FOR RECOGNITION OF NON-VERBAL VOCALISATIONS IN EXPERIMENT 3.

Chi values for recognition of non-verbal vocalisations for each emotion in Experiment 3. All analyses significant at $p < 0.0001$ level, Bonferroni corrected for 9 comparisons. Degrees of freedom for all analyses = 8.

Emotion	Chi
Achievement/Triumph	1155.2
Amusement	1357.3
Anger	1310.4
Disgust	1530.5
Fear	1097.6
Pleasure	894.0
Relief	1348.4
Sadness	894.3
Surprise	1244.5

D. CHI VALUES FOR EACH EMOTION IN EXPERIMENT 4.

Chi values for each emotion in Experiment 4. All chi analyses significant at $p < 0.0001$ level, Bonferroni corrected for 10 comparisons. Degrees of freedom for all analyses = 9.

Emotion	Chi
Achievement/Triumph	829.5
Amusement	1647.9
Anger	648.4
Contentment	489.8
Disgust	1592.7
Fear	677.5
Pleasure	964.2
Relief	868.8
Sadness	449.1
Surprise	600.7

E. CHI VALUES FOR RECOGNITION OF ACOUSTICALLY MANIPULATED EMOTIONAL VOCALISATIONS.

Chi values for recognition of acoustically manipulated emotional vocalisations. Significance levels Bonferroni corrected for 40 comparisons. Degrees of freedom for all analyses = 9.

Emotion	Stimulus type (manipulation)			
	1-channel vocoded	6-channel vocoded	Original	Rotated
Achievement	17.5	15.8	411.4§	30.7*
Amusement	299.1§	321.9§	516.3§	632.1§
Anger	142.2§	143.4§	441.9§	56.8§
Contentment	27.8*	32.8†	307.6§	40.8‡
Disgust	40.5‡	64.8§	765.0§	169.9§
Fear	21.2	36.3†	562.2§	34.4†
Relief	32.3†	49.7§	706.7§	95.8§
Pleasure	23.35	51.6§	502.2§	57.0§
Sadness	73.9§	173.9§	669.0§	186.9§
Surprise	56.1§	89.9§	669.0§	216.8§

* Indicates $p < .05$

† Indicates $p < .01$

‡ Indicates $p < .001$

§ Indicates $p < .0001$

F. ACOUSTIC ANALYSIS OF NON-VERBAL EMOTIONAL SOUNDS.

Acoustic analysis of non-verbal emotional sounds, as man per category. Note: Dur = Duration, Ampstd = Amplitude standard deviation, Ampons = Amplitude onsets, Int = Intensity, Pimin = Pitch minimum, Pimax = Pitch maximum, Pimean = Pitch mean, Pistd = Pitch standard deviation, Spcog = Spectral centre of gravity, Spstd = Spectral standard deviation. Units explained in text. Ach = Achievement/triumph, Amu = Amusement, Ang = Anger, Con = Contentment, Dis = Disgust, Ple = Sensual pleasure, Rel = Relief, Sad = Sadness, Surp = Surprise.

Acoustic measure				Stimulus type (emotion)							
	Ach	Amu	Ang	Con	Dis	Fear	Ple	Rel	Sad	Surp	Ave
Dur	1.0	1.6	1.1	1.5	0.7	1.0	1.2	0.8	1.4	0.3	1.1
Ampstd	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.06	0.0	0.1	0.1
Ampons	1.5	5.4	1.8	2.6	2.2	2.8	1.6	1.2	2.8	1.0	2.3
Int	73.5	64.9	70.1	69.2	68.3	71.4	70.7	69.7	66.5	68.9	69.3
Pimin	166.	220.6	131.6	99.7	148.3	311.4	128.1	174.5	180.4	278.9	184.6
Pimax	532.3	693.1	510.6	533.1	675.3	590.4	545.1	400.8	442.5	528.4	545.5
Pimean	415.9	359.4	283.2	219.2	352.0	443.9	243.1	283.0	274.3	395.6	327.4
Pistd	95.3	122.3	102.1	134.6	186.9	68.0	125.3	78.4	66.7	92.5	107.3
Spcog	859.4	748.2	1088.7	461.4	784.1	976.7	311.6	482.5	446.0	793.4	691.2
Spstd	491.5	879.5	550.7	556.4	630.0	408.6	327.3	752.4	319.8	521.8	543.7

G. EMOTION STORIES USED WITH HIMBA PARTICIPANTS.

Emotion	Scenario
Achievement/Triumph	You, alone, kill a jaguar without the use of a weapon
Amusement	You are being tickled and find it really funny
Anger	Man scenario: Someone sleeps with your wife
	Woman scenario: Your child has lost cattle by carelessness
Disgust	You put your hand in feces
Fear	You suddenly encounter a jaguar while alone and without a weapon
Pleasure	Your boyfriend/girlfriend/husband/wife is touching you in a sensual way
Relief	You thought you had lost your cattle but find them again
Sadness	You find out that someone close to you has died
Surprise	You come home and find that there is a rock in your hut that you didn't put there

H. CHI VALUES FOR EMOTIONS IN HIMBA PARTICIPANTS MATCHING EMOTIONAL SOUNDS TO STORIES.

Chi values for each emotion in Experiment 7, Himba participants matching emotional sounds to stories. Degrees of freedom for all analyses = 1.

Emotion	Chi
Achievement	9.8*
Amusement	53.3‡
Anger	12.5†
Disgust	20.3‡
Fear	14.7†
Pleasure	13.8†
Relief	24.8‡
Sadness	13.8†
Surprise	2.4

* Indicates $p < .05$

† Indicates $p < .01$

‡ Indicates $p < .001$

I. WESTERN PARTICIPANTS' RECOGNITION OF STIMULI PRODUCED BY HIMBA AND WESTERN POSERS.

Performance of Western participants in the recognition of stimuli produced by Himba and Western posers in Experiment 9. Degrees of freedom for all analyses = 1, 38. Ach = Achievement/triumph, Amu = Amusement, Ang = Anger, Dis = Disgust, Ple = Sensual pleasure, Rel = Relief, Sad = Sadness, Surp = Surprise.

Emotion	Himba stimuli	Western stimuli	Total	T	P-level
Ach	0.5	0.8	0.7	9.6	<0.05
Amu	0.9	0.9	0.9	1.3	N.S.
Ang	0.6	0.9	0.8	13.2	<0.001
Dis	0.8	0.9	0.9	8.1	<0.01
Fear	0.5	0.8	0.7	13.8	<0.001
Ple	0.4	0.7	0.6	21.3	<0.0001
Rel	0.9	0.9	0.9	0.6	N.S.
Sad	0.5	0.7	0.6	23.4	<0.0001
Surp	0.5	0.9	0.7	40.7	<0.0001
Total	0.6	0.7	0.7	56.1	<0.0001

J. CHI VALUES FOR RECOGNITION OF EMOTIONAL SPEECH STIMULI.

Chi values for recognition of emotional speech in Experiment 10. All chi analyses significant at p < 0.0001, Bonferroni corrected for 11 comparisons. Degrees of freedom for all analyses = 10.

Emotion	Chi
Achievement/Triumph	742.6
Amusement	1506.38
Anger	1831.1
Contentment	1031.5
Disgust	701.6
Fear	1382.7
Happiness	384.6
Pleasure	641.1
Relief	661.2
Sadness	2023.6
Surprise	1690.1

K. KAPPA SCORES FOR EXPERIMENTS 2, 3 AND 10.

Emotion	Experiment		
	Experiment 10	Experiment 2	Experiment 3
Achievement/Triumph	0.51	0.74	0.80
Amusement	0.81	0.77	0.87
Anger	0.89	0.62	0.85
Contentment	0.66	0.40	N/A
Disgust	0.52	0.93	0.93
Fear	0.77	0.59	0.78
Pleasure	0.51	0.61	0.70
Relief	0.51	0.84	0.88
Sadness	0.94	0.66	0.70
Surprise	0.86	0.49	0.84

L. CHI VALUES FOR RECOGNITION OF ACOUSTICALLY MANIPULATED EMOTIONAL SPEECH.

Chi values for recognition of acoustically manipulated emotional speech in Experiment 11. Bonferroni corrected for 44 comparisons. Degrees of freedom for all analyses = 1

Emotion	Stimulus type (manipulation)			
	Original	One-channel	Six-channel	Rotated
Achievement	192.7§	0.1	7.6	1.9
Amusement	277.7§	1.2	14.4†	14.4†
Anger	494.2§	58.1§	244.0§	4.23
Contentment	94.3§	147.5§	4.2	43.3§
Disgust	147.5§	0.4	7.6	11.9*
Fear	173.9§	34.6§	23.4§	0.4
Happy	4.2	6.1	3.1	1.2
Pleasure	30.6§	0.4	7.9	0.5
Relief	233.3§	0.1	30.6§	5.8
Sad	811.8§	5.8	63.5§	58.1§
Surprise	478.9§	2.0	2.9	1.0

* Indicates $p < .05$

† Indicates $p < .01$

§ Indicates $p < .0001$

M. CHI-SQUARE VALUES FOR RECOGNITION OF EMOTIONAL SPEECH WITH VARYING NUMBERS OF CHANNELS.

Chi values for recognition of emotional speech with varying numbers of channels in Experiment 12. Bonferroni corrected for 100 comparisons. Analyses yielding significant results in bold. Degrees of freedom for all analyses = 1. Ach = Achievement/triumph, Amu = Amusement, Ang = Anger, Con = Contentment, Dis = Disgust, Ple = Sensual pleasure, Rel = Relief, Sad = Sadness, Surp = Surprise.

Emotion	Number of channels									
	1	4	6	8	12	16	20	24	28	32
Ach	1.7	6.7	**15.0**	**36.3**	**125.2**	**60.**	**201.7**	**135.0**	**98.1**	**115.7**
Amu	4.6	1.7	1.7	3.0	6.7	9.1	**145.2**	**81.7**	**26.7**	**47.4**
Ang	**47.**	**155.7**	**89.**	**47.4**	**98.0**	6.7	3.0	**47.**	**15.0**	**31.3**
Con	1.7	**18.5**	0.7	3.0	**60.0**	3.0	**26.7**	**36.3**	**47.4**	4.63
Dis	0.2	3.0	1.7	3.0	1.7	0.2	6.7	0.7	6.7	0.7
Fear	3.0	6.7	0.7	1.7	**18.5**	11.9	6.7	11.9	**47.4**	36.30
Ple	4.6	0.7	0.2	4.6	0.7	1.7	0.7	4.6	3.0	0.00
Rel	3.0	0.7	1.7	3.0	**36.3**	**18.5**	3.0	**31.3**	**41.7**	**26.7**
Sad	4.6	11.9	**145.2**	**125.2**	**145.2**	226.9	214.1	311.3	296.3	**342.4**
Surp	0.7	**22.4**	**106.7**	**115.7**	240.0	**240.0**	**391.9**	**201.7**	**201.7**	**311.3**