Vocal and Musical Emotion Recognition: Development, Individual Differences, and Links to Broader Socio-Emotional Dimensions

Matthew C. Scott

Thesis submitted for the Degree of Doctor of Philosophy Cardiff University

January 2025

Summary of Thesis

Vocal emotion recognition and a range of abilities relating to music are positively related to positive developmental outcomes. These two forms of audio are also similar in their capacity to elicit and express emotions. However, little is known regarding similarities and differences in the recognition of emotions in these forms of audio stimuli and how these abilities develop, while possible explanations for diverging and converging emotion recognition patterns are underexplored. Further, it is currently unclear whether emotion recognition for musical and vocal stimuli relate in a similar fashion to broader socio-emotional difficulties in children. Greater understanding of these factors could contribute to a cross-condition model of audio emotion recognition development – informing future music-based interventions for children.

Chapters 2 and 3 of the present thesis examined emotion recognition for vocal prosody, instrumental and singing stimuli in adults and typically developing (TD) children, while exploring possible stimulus-level factors that may partially explain patterns of perception. Specifically, Chapter 2 explored adults' emotion recognition patterns for these audio stimulus-types, as well as their perceptions of fundamental affective dimensions arousal and valence. Stimulus acoustic features were analysed in relation to these perceptual measures. Chapter 3 focused on audio emotion recognition in TD children, assessing the possible role of children's understanding of arousal and valence dimensions during emotion recognition development. Chapter 4 shifted focus to individual differences in children's audio emotion recognition accuracy, considering emotion language comprehension as a possible condition-general mechanism of emotion recognition development. Finally, Chapter 5 examined similarities and differences between audio conditions in the relationships between

emotion recognition and socio-emotional dimensions of internalising and externalising difficulties.

Findings from all chapters indicated associations between musical vocal stimuli in relation to overall emotion recognition accuracy for adults and children. Findings in Chapter 4 suggested that these associations may be partially underpinned by a condition-general mechanism in emotion language comprehension. However, chapters 2 and 3 highlighted some condition-specific patterns of emotion recognition accuracy for certain emotions, while acoustic features did not consistently relate to specific emotions between conditions. While adults' perceptions of arousal and valence generally aligned between conditions in Chapter 2, the roles of these fundamental dimensions in children's emotion recognition appeared to vary with age and between conditions. Importantly, differences between conditions also translated to condition-specific associations with broader socio-emotional difficulties, with instrumental emotion recognition the only dimension associated with externalising difficulties in children. Findings from this thesis were integrated to inform a cross-condition model of audio emotion recognition. This model could provide a strong basis for future research on cross-condition audio emotion recognition development and inform music-based interventions.

Acknowledgements

I'd first like to thank both of my PhD supervisors – Professor Stephanie van Goozen and Dr Catherine Jones. Your belief in me was resolute, even in moments when I struggled to believe in myself. I'll be forever grateful for all the support, guidance and patience. I'd also like to extend my thanks to Dr Michael Lewis and Dr Amy Paine for their expertise and advise at various stages of this journey, and to the Economic and Social Research Council for funding this PhD.

I'm also extremely grateful for everyone who helped with data collection, and generally made my time in the Neurodevelopment Assessment Unit enjoyable and rewarding. Kate, Steve, Cass, Claire, Ellie, Alex, Dolapo, Dasha, Martha, Ellie, Amber, Carys, Taryn, Elan, and Amy – thank you all so much. Also, to everyone from my PhD year-group, it's been invaluable to share in the struggles and successes with you all, and I wish you all the very best for the future.

No one has felt the ups and downs of this experience as much as my partner Jaz. You have been patient, kind, and just generally incredible. I'll be forever grateful for you. Also, to my dog Millhouse, you may not have had a clue what was going on, but you knew I needed cuddles – thank you. Thank you to my family, particularly my parents, for supporting me unwaveringly and pre-emptively referring to me as Dr Scott.

Special recognition must go to all the children, families, and schools for taking part. They made my research experience extremely rewarding and without their willingness to help this research would not have been possible.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADHD = Attention deficit hyperactivity disorder

BF = Bayes Factor

BPVS = British Vocabulary Picture Scale

CBCL = Child Behaviour Checklist

CD = Conduct Disorder

DLD = Developmental Language Disorder

DSM-5 = Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition

GLMM = Generalised linear mixed model

LMM = Linear mixed model

NDAU = Neurodevelopment Assessment Unit

OR = Odds ratio

RAVDESS = Ryerson Audio-Visual Database of Emotional Speech and Song

SDQ = Strengths and Difficulties Questionnaire

TD = Typically developing

1. General Introduction

1.1 Overview

Music is everywhere. In fact, it exists in every known human culture (McDermott & Hauser, 2005). Much research has attributed this universality to functional links to the evolution of language, referencing overlap between the structural, cognitive, neurophysiological, and perceptual components of music and language (Fedorenko et al., 2009; Jentschke, 2016; Patel, 2010). Music is also closely related to speech, particularly vocal prosody (suprasegmental aspects of speech including intonation, stress, and rhythm), via their parallel capacity to elicit and convey emotion (Eerola et al., 2013; Day & Thompson, 2024; Juslin & Sloboda, 2011).[1] This has been attributed to their shared reliance on variations in acoustic features to express emotions (Eerola et al., 2013; Grandjean et al., 2006; Juslin & Laukka, 2003; Scherer et al., 2015).

These commonalities translate to alignment in typical emotion recognition development trajectories for musical and vocal stimuli (Heaton & Allgood, 2015; Vidas et al., 2020), which may be underpinned by shared processing mechanisms at the acoustic-perceptual level (e.g., Vigl et al., 2024). The ability to recognise emotions in others is key to children's developing social competence (Denham, 1998), while vocal emotion recognition ability, specifically, relates to a range of socioemotional dimensions in children (e.g., Chronaki et al., 2015a; Neves et al., 2021). Concurrently, children's early musical experiences, including early infant-directed singing, are key to a multitude of positive developmental outcomes (Politimou et al., 2018). Further, musical interventions and teaching relate to improvements in some of these outcomes for children (Blasco-Magraner et al.,

[1] Prosody is sometimes considered in terms of linguistic prosody (to stress/specify linguistic information) and affective prosody (to reflect emotional state of the speaker). All references to prosody in this thesis relate to affective prosody.

2021). However, few studies have examined musical and vocal emotion recognition development in tandem. Of those that have, none have attempted to explain similarities and differences between audio conditions via reference to stimulus-level features and socio-emotional developmental mechanisms, while generally involving only instrumental musical stimuli. Further, little research within either audio condition has included children with higher levels of socio-emotional or behavioural difficulties, nor examined links between emotion recognition and broader socio-emotional dimensions in these children. This limits applicability of any cross-condition model beyond typical development. Accordingly, the ultimate aim of this thesis was to contribute to a better explicated cross-condition model of audio emotion recognition development that accounts for differing forms of musical stimuli and considers individual differences in developmental processes. This could contribute to better evidenced musical applications/interventions centred on emotion recognition and socio-emotional outcomes more broadly.

This introductory chapter begins with a critical discussion of emotion theory and emotion development, before exploring links between musical and vocal emotion expression through an evolutionary functionalist lens. Focus then shifts to audio emotion processing mechanisms, including a discussion of emotion recognition development for musical and vocal prosody stimuli. Individual differences in emotion recognition are then considered, with reference to language development as a possible cross-condition mechanism of audio emotion recognition development. Finally, possible links between audio emotion recognition and broader domains of socio-emotional development are discussed, before an overview of the thesis, its core research questions, and hypotheses, are outlined.

1.2 What Are Emotions?

1.2.1 Discrete and Dimensional Theories of Emotion

Broadly, emotion theories can be bisected into discrete and dimensional perspectives. Discrete theories treat emotion categories as innate and distinct – both neurobiologically and in terms of their causative relationship with cognitions and expressive behaviours (for review, see Keltner et al., 2019). Although variation in the subjective experience of emotions is accepted, this is generally thought to stem from functionally separable cognitive evaluations, while language describes, rather than interacts with or constitutes, emotion perceptions/experience (Ekman & Cordano, 2011; Izard, 2009).[2] Some discrete theories place greater emphasis on interaction between cognition/language and emotion as key to modulating the expression and perception of emotions, at various stages of processing (e.g., Ellsworth, 2013; Moors et al., 2013). Conversely, dimensional perspectives reject the notion of distinct neurobiological correlates for specific emotions (Barrett, 2013; Russell, 2003). Instead, they centralise language, via emotion concepts, as key to *constituting* emotion perceptions or experiences, built from the context-dependent interpretation of more fundamental affective dimensions – most commonly arousal (low to high energy) and valence (negative to positive – e.g., Barrett, 2017; Posner et al., 2005; Russell, 2003). However, the extent to which emotion categories are mapped onto these valence and arousal dimensions consistently across individuals and cultures varies greatly between theories. [3]

Meta-analytic neuroimaging evidence substantiates both discrete and dimensional perspectives (Lindquist et al., 2012; Vytal & Hammond, 2010). While methodological

[2] 'Emotion perception' is used in the present thesis to refer to an inferred emotion that may match or differ from the expressive intention, while 'emotion recognition' refers to alignment between the expressive intention and inference.

[3] This is a simplification, and emotion theories vary greatly along this spectrum relying on varying degrees of appraisal. For a recent review, see Scherer (2022).

differences may underpin this (Celeghin et al., 2017), ultimately, the importance of discrete emotion categories to the conscious experience and perception of emotion is unequivocal, even within frameworks centralising arousal and valence within these processes (Cameron et al., 2015). Recent reviews have echoed this sentiment, stressing convergence between discrete and dimensional perspectives and the need to consider both emotion categories and more basic affective features within research (Harmon-Jones et al., 2017; Scherer, 2022). Woodward et al. (2021) offered support for this convergence – finding that while children are inclined to categorise vocal stimuli, category boundaries also shift based on the overall expressive arousal of the speaker. They concluded that learning about the perceptual features associated with specific emotions, as stressed within discrete theories and often overlooked within dimensional ones (Keltner et al., 2019), is key to emotion development. However, the dynamic interactions between this learning process and the development of language-mediated conceptual understanding of emotions, as stressed within dimensional theories, is also fundamental to perceptual emotion development (Woodward et al., 2021). Indeed, given activation in brain regions previously attributed to semantic processing during emotion experience, it appears likely that language at the very least interacts with, if not partially constitutes, emotions (Satpute & Lindquist, 2021).[4]

Importantly, the perceptual categorisation of emotion expressions allows individuals to comprehend their meaning (Ross & Spalding, 1994) – aiding decision-making (Markman & Rein, 2013), facilitating coping mechanisms (Zimmer-Gembeck et al., 2013), and encouraging the formation and maintenance of friendships (Lemerise & Arsenio, 2000). Further, as will be discussed, the functional ability to recognise emotions within these categorical bounds is related to a range of positive developmental outcomes for children (e.g.,

[4] The term 'dimensional' was used to reflect terminology used in music research. However, these are often referred to more generally as 'constructionist' theories – see Warrenburg (2019) for discussion.

Chronaki et al., 2015a; Cooper et al., 2020).[5] Accordingly, the present thesis focuses on emotion recognition as a functional developmental ability, but also considers the possible role of underpinning arousal and valence properties in how emotions are discriminated, and in the development of this ability. A focus on functionality also facilitates further consideration of the development of emotional abilities.

1.2.2 Typical and Atypical Emotional Development

A functional perspective on emotion development considers emotions as relational processes – attempts by an individual to 'establish, maintain, change, or terminate' their relation to their environment (Campos et al., 1989, p.395). This allows focus on the ways in which individuals' interactions with their environment shift across development, to meet implicit or explicit emotion-related goals (e.g., maintaining and improving wellbeing – Cole, 2016). Within this context, emotion expressions, such as via face and voice, are outcomes of these relational processes and operate functionally before they are understood or perceived as explicitly emotional (Buss et al., 2019). Indeed, preverbal infants form functional associations between certain events and vocal/facial emotions (Ruba et al., 2018), while they are also able to discriminate emotions based on their perceptual properties (Ruba & Repacholi, 2020). Later in development, language acquisition coincides with an improved ability to categorise stimuli based on emotional meaning – initially basing these categorisations on valence and arousal (Gao & Maurer, 2010; Shablack & Lindquist, 2019), before learning to attribute them to discrete emotion categories (Russel & Widen, 2002). This latter ability, for both faces and voices, develops throughout childhood, and in the case of vocal expressions, into adolescence (Grosbras et al., 2018; Widen, 2013). Developmental difficulties, and associated difficulties recognising emotions, can arise via atypicality within

[5] This thesis adopts the term 'emotion recognition' in line with past research in the field, but this can also be considered as 'emotion reasoning' or 'emotion inference', depending on theoretical orientation (see Ruba & Pollack, 2020).

these complex interactions with the environment, modulated by psychobiological vulnerability (Eme, 2017). Within this framework, emotions can be considered in terms of their internal and external functionality (e.g., Oatley, 1992). Internal functionality relates to emotions and their interaction with changes in internal goals, while external functionality refers to how expressions of emotion convey information for social interaction (Juslin, 2018). Through an evolutionary lens, this dual functionality can highlight the ways in which music may fit into this developmental picture and underline its commonalities with vocal emotion expressions.

1.3 Emotion Expression in Music and Voice - An Evolutionary Functionalist Perspective

1.3.1 What Does Music Have to do With Emotions?

As noted, instrumental music reliably conveys and elicits emotions (Juslin & Sloboda, 2011). Indeed, of 141 Western adults asked what music expresses, 100% provided emotion as one of their answers, while cross-cultural evidence suggests universal recognition of some musical emotions (Argstatter, 2016; Laukka et al., 2013b). However, certain expressive features are utilised and perceived differently across cultures (Athansopooulos et al., 2021; Laukka et al., 2013b), while studies with infants indicate that various aspects of musical development, such as rhythm discrimination, are strongly influenced by culture (Stewart & Walsh, 2005). Accordingly, music's capacity to convey and elicit emotions can be attributed to (often universal) structural psychophysical cues such as tempo, pitch patterns, and loudness (Gabriellson & Lindstrom, 2010; Juslin & Sloboda, 2011), within more culture-dependent musical conventions that convey further symbolic meaning (e.g., scales, ragas - Balkwill & Thompson, 1999).

Despite music's apparent emotional resonance, some have argued for a firm distinction between music and other forms of emotion expression on account of music only

expressing and inducing emotion symbolically, rather than with any innate social function or consequence (for discussion, see Masatsaka, 2009). For cognitive psychologist Steven Pinker (1997), music is simply 'auditory cheesecake' – a pleasurable but useless by-product of language development. There are various lines of evidence that question this, however. For example, instrumental music can communicate socially relevant semantic information in a way analogous to language (Fritz et al., 2019; Koelsch et al., 2004), while musical and social emotion understanding also appear to draw on a shared capacity to represent emotions from social cues (Siu & Cheung, 2017). Considered alongside music's cross-cultural/historical ubiquity and its powerful emotional impacts, this suggests that music does have adaptive significance, in line with other forms of emotion expression (Trehub, 2003).

This general emotional capacity of music has been partially attributed to its functional evolutionary origins. Internally, some evidence suggests that many evolved affective mechanisms function in response to music similarly to other auditory stimuli (Juslin, 2018). For example, reflexive responses in the brainstem that transmit sensory information between the body and rostral brain structures (Venkatram et al., 2017) operate in response to both music (Harmat et al., 2014) and acoustic stimuli more generally (Simons, 1996) - modulating salience of emotional input and increasing arousal. Further, the internal mirroring of vocal emotion expressions via pre-motor brain regions to facilitate emotion contagion (Herrando & Constantinides, 2021) also operate in response to music (Koelsch, 2014). Similarly, evolved external functions of emotions, including communicating meaning to aid social organisation (Buck, 2014), may be relevant to music. Indeed, some have referenced overlapping neurophysiological, perceptual, and cognitive correlates of music and language as indicative of music's evolved role as a tool to facilitate shared intentionality and build/maintain social cohesion (Cross, 2014; Patel, 2010). Given music's ability to express emotions, as well the

psychophysical features involved, this external functionality underlines possible commonalities between music and other forms of emotion stimuli, particularly vocal prosody.

1.3.2 A Shared Evolutionary Basis to Emotions in Music and Vocal Prosody

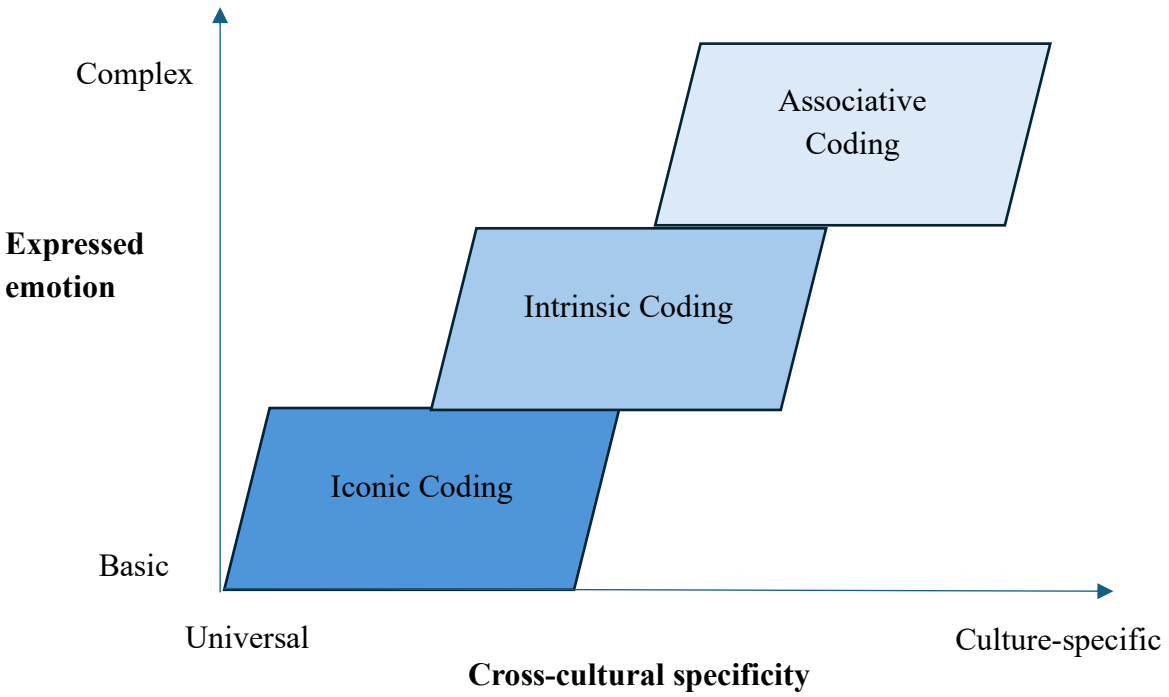
Research on vocal emotion expression and perception generally focuses on affect bursts (non-verbal expressions such as laughs and grunts) and/or prosody (Vidas et al., 2018). Prosodic cues include loudness, tempo, intonation, pitch, rhythm, and timbre (Grandjean et al., 2006), and each of these cues are also central to emotion communication in music (Eerola et al., 2013; Juslin & Laukka, 2003). This has led to a research focus on possible commonalities in musical and prosodic emotion expression – including their evolutionary origins and implications for our understanding of emotions and emotion development. One prominent theory contends that musical and prosodic emotions evolved in parallel from animalistic affect bursts (Scherer, 1995). Others have expanded on this theory, placing 'musilanguage' – a primitive, non-linguistic form of emotion expression based on intonation - as a necessary intermediary between affect bursts and the evolution of emotional expression in music, speech, and language (Brown, 2017). Importantly, Clark et al. (2015) hypothesise that these early expressions had adaptive significance as a means to 'mentally rehearse' affectively significant social routines (e.g., infant bonding), without the potential physical, neural, or behavioural costs. Developmentally, these evolving music-like interactions may be reflected in the acoustically exaggerated, music-like nature of early infant-mother interactions (Saint-Georges et al., 2013). In each case, evolved parallels between music and prosodic/intonational aspects of speech, via acoustic features, are emphasized.

1.3.3 How Does Music Express Emotion?

One model of musical emotion expression comes from Juslin (2013). They conceptualise musical emotion expression (predominantly in instrumental forms) via

differing levels of hierarchically organised 'coding', termed iconic, intrinsic, and associative (Figure 1.1). Iconic coding represents evolved, functional links between musical and vocal emotion expression based primarily on shared acoustic-perceptual mechanisms. The intrinsic level adds complexity to emotion expressions via syntactical elements within the music itself, such as the use of dissonant or consonant harmonic progressions to manipulate tension, arousal, and other affective dimensions. The associative level denotes individual associations between aspects of the musical excerpt and personal experiences/objects. As Figure 1.1 shows, moving up each coding layer of the hierarchy is associated with an increase in the breadth of emotions that can be communicated. Higher layers are also associated with greater cross-cultural differences in emotion expression and perception (Juslin, 2013). Importantly, Juslin (2013) contends that it is the bottom iconic level that most powerfully and consistently expresses emotion and explains cross-cultural recognition of many emotions in both music (Laukka et al., 2013b) and voice (Chronaki et al., 2018; Laukka et al., 2013a).

Figure 1.1 – Adapted version of Juslin's (2013; 2018) Multi-Layer Model of Musical Emotion Expression



Although a functional acoustic-perceptual link between music and voice is widely accepted, there is disagreement regarding the extent to which acoustic patterns align more closely with discrete emotions, or affective dimensions such as valence and arousal (Warrenburg, 2019). As discussed, this reflects debates within emotion research more generally (Scherer, 2022).

1.3.4 Musical and Vocal Emotion Expression – Linked at the Level of Emotions or Dimensions?

Some researchers posit that the functional link between musical and vocal emotion expression relates to shared expressive patterns for specific emotions. For example, building on the iconic level of Figure 1.1, music has been conceptualised as a 'super-expressive voice' – an acoustically exaggerated counterpart, with shared expressive features grouped within emotion categories (Juslin & Sloboda, 2011; Scherer, 1995). This mirrors the general predictions made within some discrete emotion theories – that each emotion encompasses distinct physiological, subjective, and expressive components (Keltner et al., 2019). In support of this theory, Juslin and Laukka's (2003) meta-analysis indicated consistency between music and voice in terms of the acoustic cues communicating emotions happiness, sadness, anger, fear, and tenderness. In further support, overall musical and vocal emotion recognition ability are significantly correlated in adults (Laukka & Juslin, 2007) and typically developing (TD) children (Heaton & Allgood, 2015; Vidas et al., 2018). Further, individuals with congenital amusia (impaired music processing despite normal peripheral auditory processing, cognitive functioning, and music exposure) have difficulties recognising emotions in vocal prosody (Lima et al., 2016; Thompson et al., 2012), implying general mechanisms underpinning emotion recognition for music and voice. This perspective does not implicitly deny the existence of arousal and valence dimensions. Indeed, Juslin and Lindstrom (2010) propose a model, shown in Figure 1.2 below, within which the most

observed musical emotions are mapped onto valence and arousal dimensions, including proposed shared acoustic correlates for musical and vocal expressions (Juslin & Lindstrom, 2010; Juslin & Laukka, 2003) [6]. However, as Figure 1.2 shows, these acoustic features are tied to emotion categories, while arousal and valence are considered secondary dimensions that can capture certain aspects of emotion perception or experience (Juslin, 2018).

[6] The original model referred to tenderness. However, calmness and tenderness have similar acoustic correlates (Lindstrom, 2006; Juslin, 1997) and tender music is often perceived as calmness (Micallef Grimaud & Eerola, 2022)

Figure 1.2 - Expressed Emotions Categorised Along Independent Planes of Valence and Arousal, With Associated Acoustic Features (adapted from Juslin & Lindstrom, 2010)

Contrastingly, dimensional theories of musical and vocal emotion expression claim that rather than being secondary affective features, arousal and valence dimensions are fundamental to *substantiating* emotion experience and perceptions (Cespedes-Guevara & Eerola, 2018). Notably, they dispute the direct mapping of expressive acoustic information to

discrete emotions, as in Figure 1.2 above. Instead, emotion recognition involves the categorisation of variations in expressed arousal and valence, via language-mediated conceptual understanding of emotions (Barrett, 2017; Russell, 2003). Consequently, emotion recognition patterns may be more variable. Indeed, while there are some similarities between music and voice in adult's emotion recognition accuracy for specific emotions (Laukka et al., 2013a; Laukka et al., 2013b), differences also exist. For example, anger is salient in high intensity prosodic emotions (Elfenbein et al., 2021; Sauter et al., 2013), but often difficult to recognise in music (Micallef Grimaud & Eerola, 2022; Vidas et al., 2018). Conversely, while happiness is recognisable in music from an early age (Dalla Bella et al., 2001; Vidas et al., 2018), difficulties recognising prosodic happiness can persist through adolescence (Chronaki et al., 2018) and into adulthood (Lausen & Hammerschmidt, 2020).

Within this dimensional perspective, these discrepancies arise because acoustic features such as sound level and tempo/speech rate vary along arousal and valence dimensions, rather than being mapped atop them within a discrete emotion category (as in Figure 1.2). Indeed, developmentally, TD children's ability to distinguish musical sadness and happiness develops at the same stage as musical entrainment – the synchronisation to musical rhythms via autonomic, motor, and/or perceptual systems (Dalla Bella et al., 2001). For Cespedes-Guevara and Eerola (2018), this centralises variations in tempo and the associated expression of arousal as key to the early ability to distinguish musical happiness and sadness. They also claim that it is at this more basic dimensional level of expression that musical and vocal emotions are linked.

1.3.5 An Integrated Theoretical Perspective on Musical and Vocal Expression That Centralises Arousal

For the present thesis, the apparent importance of both discrete emotions and dimensions for musical and vocal emotion expression supports an integrated consideration of

each these affective components. Indeed, perceptions for arousal and valence dimensions can account for a high amount of variance in emotion ratings for vocal stimuli (Sauter et al., 2010) and music stimuli (Eerola & Vuoskoski, 2011), suggesting a high degree of alignment between each aspect of emotion. However, for music stimuli, Eerola & Vuoskoski (2011) also found that perceptions of arousal and valence dimensions were more reliable for ambiguous stimuli relative to emotion judgements. This may imply primacy for these affective dimensions - supporting the dimensional perspectives' focus on arousal and valence as fundamental features that substantiate emotion perceptions (Cespedes-Guevara & Eerola, 2018).

However, further evidence points to complexity in terms of the relative importance of arousal compared to valence for audio emotion expressions. Indeed, there may be a more integral role for expressed arousal, relative to valence, in vocal and musical emotion expression. This diverges from some dimensional theories, within which emotions emerge from a 2-factor cognitive interpretation of valence and arousal (Barett & Bliss-Moreau, 2009; Russell, 2003), and from recent facial expression research stressing a greater role for valence, relative to arousal, during emotion perceptions (Woodward et al., 2022). For example, while arousal level is accurately perceived at any level of expressed emotion intensity in vocal stimuli, perceptions of valence and specific emotions are significantly less accurate at both minimal and maximal levels of expressed intensity (Holz et al., 2021). This may indicate that high arousal, high intensity vocal signals are processed automatically (possibly via hardwired brain responses such as brainstem reflexes – Gomez-canon et al., 2021), while a decrease in intensity allows a more fine-grained analysis of 'affective semantics' involving valence and emotion category (Holz et al., 2021). This aligns with the proposed relative importance of arousal within musical emotion expression, possibly drawing on similar automatic processing mechanisms (Cespedes-Guevara & Eerola, 2018). Indeed, acoustic features such as loudness

and tempo are positively associated with perceived arousal in both music and voice, while arousal is more strongly predicted by acoustic features compared to valence in both audio conditions (Bänziger et al., 2015; Llie & Thompson, 2006; Coutinho & Dibben, 2013).

Evidence also implies a need to integrate elements of discrete theories of audio emotion expression – namely, their assertion that acoustic properties can communicate emotions directly (Juslin, 2013). Indeed, although arousal is strongly related to emotion communication via acoustic features in the voice, there are also 'unique combinations of distal cues and proximal percepts carrying information about specific emotion families, independent of arousal', particularly for emotions such as anger (Bänziger et al., 2015, para. 1). Further, some prosody emotion recognition research indicated greater cross-cultural preservation of acoustic correlates of vocal emotion categories, relative to both valence and arousal (Cowen et al., 2019). Similarly, for music, research suggests that certain acoustic features can communicate arousal *and* certain emotions across different cultures (Athansopolous et al., 2021). This would appear to support elements of the multi-layered model in Figure 1.1 in terms of evolved, universal features of musical emotion expressions, possibly due to functional links to the voice. However, it also appears that these links are tied at a fundamental level to *both* discrete emotions and dimensions such as arousal. This is explainable via recent neuroimaging research, that found separable but overlapping neural processing systems for arousal/valence and emotion categories during vocal emotion perceptions (Giordano et al., 2021). This may suggest a reconciliation of previously opposing dimensional and discrete theories of audio emotion expression, while explaining more direct links between expressive features and emotion perceptions for musical and vocal stimuli than assumed within dimensional theories (Bänziger et al., 2015; Athansopolous et al., 2021).

While the exact nature of the interaction between these expressive and perceptual systems is beyond the scope of the present thesis, the ability to recognise stimuli as discrete

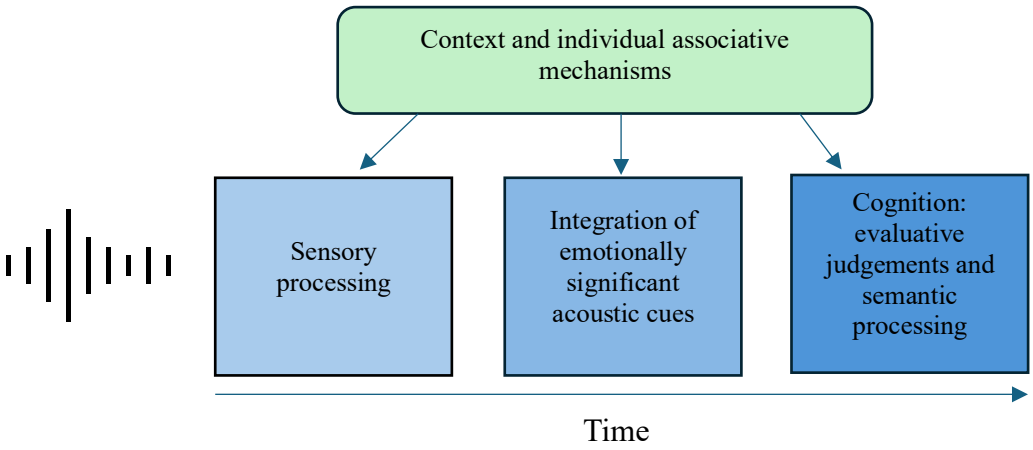
emotions has important implications for positive developmental outcomes (Chronaki et al., 2015a; Cooper et al., 2020). However, the present thesis assumes that the affective dimensions of arousal and valence represent independent dimensions that also have a role in substantiating discrete emotion perceptions. As will be discussed, this is particularly important for developmental research, given the interplay between perceptual understanding of valence and arousal, emotion categories, and language across development (Ruba et al., 2018; Woodward et al., 2021; Widen 2008).

1.4 Audio Emotion Perceptual Processing Mechanisms

1.4.1 Audio Emotion Recognition – A Multi-Stage Process

Based on a review of neuroimaging evidence, Schirmer and Kotz (2006) outlined a 3-stage model for the recognition of emotion prosody. This involves the initial low-level sensory processing of input, before an integration of emotionally significant acoustic information and a final higher-order cognitive interpretation of emotion expression. Context and individual associative mechanisms are thought to affect processing throughout. A simplified depiction of this process is outlined in Figure 1.3. Bestelmeyer et al. (2014) offered broad support for this model, isolating an initial process involving the decoding/integration of acoustic information within primary auditory areas and the amygdala, and a later process involving the cognitive interpretation of emotional meaning within prefrontal and anterior insula regions (Bestelmeyer et al., 2014).

Figure 1.3 – Adaptation of Schirmer and Kotz's (2006) Multi-Stage Model of Emotion Prosody Processing



This multi-stage model also appears relevant for music. Indeed, neurophysiological research indicates extensive overlap during emotion processing of musical and vocal stimuli (Escoffier et al., 2013; Frühholz et al., 2016; Paquette et al., 2018; Proverbio & Piotti 2022). Importantly, in alignment with vocal research, overlap is apparent during acoustic integration (Peretz et al., 2015) and the cognitive interpretation of emotion stimuli (Schirmer et al., 2012). However, the extent to which stages of this model represent separable processes organised linearly, or have a more interactive relationship, will differ between theories. Specifically, those that propound a constitutive role for language within emotion perceptions (i.e., some dimensional theories) would predict more interaction between cognitive interpretation and the low-level integration of acoustic information, compared to those that consider cognition and language as separable aspects of emotion recognition (i.e., discrete theories). Indeed, a recent review highlighted a possible fundamental role for brain regions involved in higher-level, language-based processing during the perception of low-level features of emotion stimuli (Satpute & Lindquist, 2021). Regardless, findings suggest overlap between music and voice in terms of neural processing during emotion recognition. However,

there remains some processing differences between music and voice that may have implications for research on musical and vocal emotion recognition development.

1.4.2 How Music and Voice Differ - Possible Mechanistic and Developmental Implications

While musical and vocal emotion processing may have a shared evolutionary basis and understanding each form of emotion expression has adaptive significance (Trehub, 2003), vocal expressions also have a unique role in development. Caregiver voices are integral in directing attention to other sources of social information (Bryant & Barrett, 2008; Trainor et al., 2000), while newborn infants show preference towards human voices relative to non-social auditory stimuli (Ecklund-Flores & Turkewitz, 1996). This is reflected in voice-specialised auditory processing regions, including the superior temporal sulcus/gyrus (Belin et al., 2002), while these regions are also involved in processing emotion prosody from early in development (Grossman et al., 2010). Reybrouk and Podlipniak (2019) also note that music has a general tendency to direct attention to the acoustic material itself, while speech, at least when including language, tends to refer to something external in the environment.

Evidence also suggests that there are music-specific areas of the brain involved in auditory processing. For example, emotional music relates to preferential activation in the superior temporal gyrus, while vocal prosody relates more strongly to activation in the superior temporal sulcus (Whitehead & Armony 2018). Importantly, this music-specific processing held for both instrumental *and* singing stimuli. Given acoustic similarities between the prosodic and singing stimuli, this may indicate a higher-order music-specific processing mechanism, possibly linked to a more fine-grained musical pitch processing system (Zatorre & Baum, 2012).

Processing differences may have implications for the ways in which musical and vocal emotion recognition develop. For example, the importance of vocal expressions in directing attention to other sources of social information (Trainor et al., 2000) could lead to variation in how emotion understanding for music and the voice develop. For music, difficulties with the processing of pitch, and their organisation within melodies, has direct clinical implications (Stewart et al., 2006). This may give rise to links between music-specific processing mechanisms and some aspects of socio-emotional development. Accordingly, a focus on singing, given its alignment with both prosodic and instrumental forms of expression, may illuminate similarities and differences between prosodic and instrumental forms of emotion expression. Indeed, singing involves the same melodic conventions as instrumental music (Coutinho et al., 2014) and operates similarly to prosody in drawing attention to socio-emotional information early in development (Schubert & McPherson, 2015). Pertinent to the present thesis, this early developmental role of singing may provide unique insights regarding how emotion recognition for music and voice develops.

1.4.3 Singing – A Conceptual and Developmental Bridge Between Instrumental Music and Vocal Prosody

Research on emotion expression and perception in singing is limited but indicates similarities with prosody. There are commonalities regarding the acoustic properties underpinning different emotions in prosody and singing (Scherer et al., 2015). Those related to arousal appear particularly strongly related, possibly due to the physiological constraints pertinent to the speaking and singing voice (Scherer et al., 2017). Despite these similarities, prosodic emotions are easier to recognise than sung emotions for adults (Livingstone & Russo, 2018). Singing is also uniquely tied to instrumental music via musical constraints, and understanding of these conventions is important for accurate emotion recognition (Balkwill & Thompson, 1999). For example, musical modes – specific scale variations that are often

either major or minor[7] - are important in communicating positive (major) and negative (minor) emotions in both singing and instrumental music (Eerola et al., 2013; Juslin & Laukka, 2003). Accordingly, singing may bridge some of the voice-specific and music-specific aspects of audio processing, illuminating the nature of similarities and differences in perceptual patterns.

Singing also forms a bridge between instrumental music and voice regarding the *development* of emotion recognition. Infant-directed song is more effective at drawing attention to the mouth – known to support speech and language development – compared to infant-directed speech (Alviar et al., 2023). Later in development, 4-6-year-old children are more likely to categorise ambiguous stimuli as song rather than speech, while older children do not, suggesting a shared developmental basis to the comprehension of speech and song (Vanden Bosch der Nederlanden et al., 2022). Schubert and McPherson (2015) implicate the 'music-like' quality of early mother-child interactions as fundamental to pre-linguistic emotion comprehension and the bonding of mother and child. It is suggested that the early ability to draw meaning from these early music-like interactions is biologically predisposed (Trehub & Schellenberg, 1995; Trehub & Trainor, 1998) and underpinned by developing sensitivity to acoustic features such as loudness, timing patterns, and pitch (Flom & Bahrick, 2007; Trehub, 2001). It is from this base that more advanced cognitive aspects of vocal emotion recognition are thought to develop (Boone & Cunningham, 1998). Correspondence between these early music-like interactions and the 'musilanguage' proposed as an evolutionary origin of emotion expressions in voice, music, and language, is clear (Brown, 2017). Accordingly, findings highlight evolutionary functionality to musical emotion expression and implicate singing as fundamental to the typical development of emotion

[7] There are seven modes in Western music, but these are often grouped as either major or minor in research considering emotion expression.

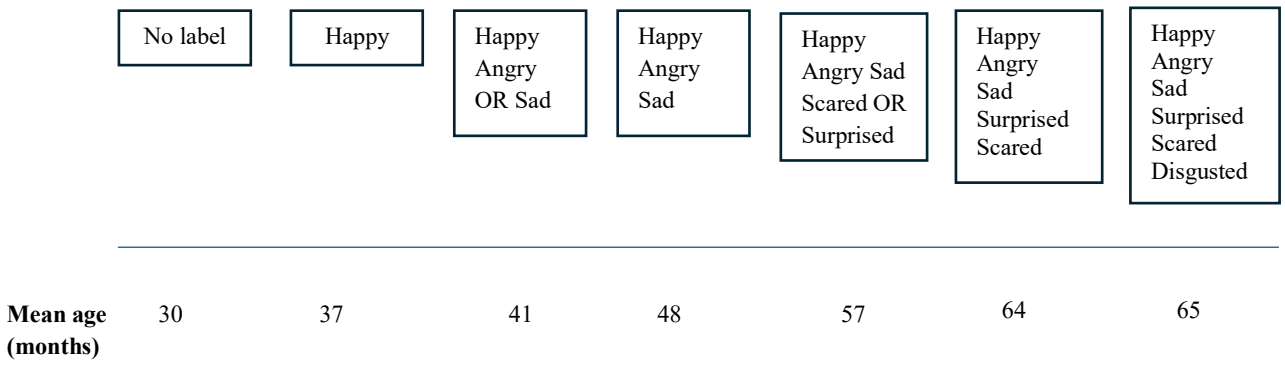
recognition in audio conditions. However, past emotion recognition research has not incorporated these three forms of audio emotion stimuli. This thesis fills this gap, to illuminate some mechanistic explanations for similarities and differences between the development of musical and vocal emotion recognition.

1.5 Typical Development of Emotion Recognition in Audio Stimuli

1.5.1 Broad-to-Differentiated Emotion Recognition Development

The development of emotion recognition begins with an initial understanding of emotion expressions' functional associations with situations/outcomes, as well an ability to distinguish them based on their perceptual properties (Ruba et al., 2018; Ruba & Repacholi, 2020). Later in development, children become more adept at categorising stimuli as discrete emotions. Based on aggregated data with over 1000 children, Widen (2013) suggests that this initially involves growing understanding of expressed valence – i.e., an ability to separate emotion expressions based on them being 'good' or 'bad' - before a more differentiated understanding of different emotions. This model is outlined below (Figure 1.4).

Figure 1.4 – Adaptation of Widen's (2013) Valence-Centred Model of Facial Emotion Recognition Development



Recent research supported this conceptualisation, finding that children and adults relied primarily on valence to freely sort facial stimuli based on perceived similarity of emotion expression, while the influence of emotion category understanding increased with age (Woodward et al., 2022). In line with an integrated theoretical perspective, sorting patterns based on valence and emotion categories were strongly related (Woodward et al., 2022). However, this model is based on facial stimuli, and whether such a pattern would be relevant for audio stimuli remains unexplored. As is discussed below, it may be that understanding of expressed arousal is more important for audio stimuli than it is for facial stimuli.

1.5.2 The Development of Emotion Recognition for Audio Stimuli

Vocal prosody emotion recognition is a key developmental skill – facilitating social interaction and positive socio-emotional adjustment (Denham, 1998; Saarni, 1999). However, compared to facial expressions, there is limited research on prosody emotion recognition development, and even less for music stimuli. While the evidence that does exist is mixed, emotion recognition patterns for both prosody and music stimuli appear partially explainable via a) differences in the stimuli adopted and b) increasing sensitivity to arousal and valence with age. Most prosody research indicates that there is rapid emotion recognition development between pre-schoolers (4-5 years) and children (6-9 years), and some continued development into adolescence (Chronaki et al., 2015b; Grosbras et al., 2018; Sauter et al., 2013). While studies adopting short prosodic ‘bursts’ as stimuli have found earlier development of happiness and sadness relative to anger and fear (Grosbras et al., 2018), those using longer stimuli have generally found that anger and sadness were the earliest emotions to develop (Nelson & Russell, 2011; Sauter et al., 2013; van Zonneveld et al., 2019; Zupan, 2015). Given some of the proposed acoustic differences between anger and fear stimuli (e.g., pitch variability - Figure 1.2), it may be that these longer stimuli allowed for interpretation of

expressive differences between anger and fear, while this may have also contributed to the greater relative accuracy for sad stimuli (e.g., via interpretation of differences in speech rate). Some researchers have suggested that general biases towards perceiving negative emotions in young children may reflect a lack of valence-based understanding and greater importance of arousal for vocal relative to facial emotion recognition, particularly early in development (Nelson & Russell, 2011). Indeed, general patterns of high accuracy for sadness early in development (Sauter et al., 2013; Grosbras et al., 2018) could also be linked to an early sensitivity to arousal, given that other commonly presented emotions (happiness, anger, fear) are hypothetically high in arousal. This would align with research highlighting a stronger link between expressive acoustic features and perceived arousal (Bäzinger et al., 2015) and greater salience of arousal properties, relative to valence, in vocal stimuli (Holz et al., 2021).

For instrumental music, research suggests early development of happiness and sadness emotion recognition, before the development of accuracy for other emotions including fear and anger (Franco et al., 2017; Kratus, 1993; Nawrot, 2003; Stacho et al., 2013). These patterns have been linked to developing sensitivity to acoustic features, and associated arousal and valence properties. For example, children's emotion judgements before 6-years are sensitive to variations in arousal-related acoustic features tempo and loudness (Dalla-Bella et al., 2001; Kragness et al., 2021; Mote, 2011), while sensitivity to musical mode develops around 6-8 years (Dalla-Bella et al., 2001). As musical mode is closely linked to perceived valence (Gomez & Danuser, 2007), this may suggest early sensitivity to expressed arousal, before later developing sensitivity to valence aligns with an increase in the emotions that can be accurately recognised. This salient valence indicator in music may also create differences in how valence is perceived between audio conditions, with implications for emotion recognition accuracy (Weninger et al., 2013).

For singing stimuli, findings appear closely aligned with instrumental music. Children as young as 3 can distinguish happiness and sadness (Franco et al., 2017), while 5-year-olds accuracy levels for these emotions matches adults (Morton & Trehub, 2007). Children's singing and instrumental music recognition patterns also appear to align when considering a wider range of emotions, such as anger and fear (Dolgin & Anderson, 1990). Again, in line with instrumental music research, 4-year-old children can expressively communicate happiness and sadness via singing, primarily through the utilisation of tempo and loudness cues (Adachi & Trehub, 1998).

Overall, changes in sensitivity to valence and arousal properties, and associated acoustic features, may be key to emotion recognition development for vocal and musical stimuli. Seemingly, arousal may be a particularly salient affective dimension, particularly early in development. This aligns with theoretical and empirical research asserting a fundamental role for arousal within vocal and musical emotion expression and perception (Cespedes-Guevara & Eerola, 2018; Holz et al., 2021). Emotion recognition patterns also point to a key developmental phase between pre-school and early childhood across stimulus types. However, there are some discrepancies in patterns between music and prosody for certain emotions, with happiness recognition seemingly earlier to develop for music relative to prosody stimuli, and anger more salient early in development for prosody relative to music stimuli (Nawrot, 2003; Sauter et al., 2013; Stacho et al., 2013). As noted, condition-specific factors such as physiological constraints and musical mode may partially underpin developmental differences between conditions. Accordingly, research considering musical and vocal stimuli in tandem may lead to unique insights into musical and vocal emotion recognition development.

1.5.3 Cross-Condition Research – Findings and Opportunities

A cross-condition approach facilitates direct consideration of the commonalities between musical and vocal emotion recognition development. Two developmental studies have taken this approach. Findings indicated parallel development for recognition of emotion in music and voice, with a significant positive correlation between instrumental and vocal (both prosody and affect bursts) emotion recognition, across different stages of development (Heaton & Allgood, 2015; Vidas et al., 2018). The most pronounced similarities appear to be between instrumental music and prosody, with recognition accuracy for prosody able to predict accuracy for instrumental music independent of age and music training (Vidas et al., 2018). However, certain condition-specific patterns were apparent. Most notably, anger was easier to recognise for prosody compared to instrumental stimuli, particularly in younger children (Vidas et al., 2018). Further, Heaton and Allgood (2015) found that musical stimuli were easier to recognise overall, supporting the idea that music acts as a highly expressive form of emotion communication (Juslin, 1997; 2018). The authors concluded that findings indicated a 'cross-condition model of auditory emotion recognition development' (Heaton & Allgood, 2015, p. 402). Considering its developmental importance (McPherson & Schubert, 2015), a cross-condition model could be better explicated by incorporating singing stimuli. Further, given the emotion recognition development patterns discussed, such a cross-condition model may be more closely linked to a heightened early sensitivity to/understanding of arousal properties, rather than the valence-based development pertinent to facial expressions (Figure 1.4).

1.6 Individual Differences in the Development of Audio Emotion Recognition

1.6.1 Individual Differences Within Different Processing Stages

As discussed, Bestelmeyer (2014) considers audio emotion recognition as a multi-stage process, including the low-level perception and integration of acoustic information, and higher-order cognitive interpretation of emotion expression. Within this framework, individual differences may arise at either stage.

Past research on vocal and musical emotion recognition has mainly focused on individual differences in the low-level acoustic-perceptual stage of this model. Individual differences in sensitivities to acoustic features (sometimes termed musical aptitude) have been proposed as central to the relationship between musical and vocal emotion recognition accuracy. For example, Vigl et al. (2024) found that individual differences in musical aptitude, mediated by the ability to perceive subtle differences in prosodic stimuli, related to vocal emotion recognition accuracy independent of years of formal music training – a general pattern supported by meta-analytic evidence including adults and children (Jansen et al., 2023). This would indicate that correlated emotion recognition accuracy between music and voice may be explainable via low-level acoustic processing similarities that operate independent of formal music training (Escoffier et al., 2013; Proverbio et al., 2020).

Given the multiple stages involved in audio emotion recognition, individual differences at the acoustic-perceptual level offer an incomplete explanation for similarities and differences in the development of musical and vocal emotion recognition accuracy. Indeed, Trimmer and Cuddy (2008) found that emotional intelligence, over and above any differences in the perception of low-level acoustic features, explained emotion recognition accuracy for instrumental music and prosody in adults. Further, emotion inferences from vocal and musical stimuli appear to tap more strongly into a general neural system of social

cognition, including superior frontal and anterior cingulate regions, than into one involved in the perception of shared acoustic features (Escoffier et al., 2013). Correspondingly, individuals with congenital amusia are impaired in not only vocal emotion recognition (Thompson et al., 2012), but also facial emotion recognition (Lima et al., 2016). For Lima et al. (2016), this points to a more general socio-emotional processing mechanism common to music and other forms of emotion expression that aligns with the cognitive process of emotional meaning interpretation in Bestelmeyer et al.'s (2014) multi-stage model. While analogous developmental research is lacking, increasingly, language development is being placed at the core of these developing cognitive aspects of emotion recognition.

1.6.2 Language as a Shared Developmental Mechanism of Audio Emotion Recognition Development

1.6.2.1 Individual Differences in Language and Emotion Recognition Development.

It has been suggested that language scaffolds the development of emotion knowledge – shaping the way in which emotions are perceived and understood (Shablack & Lindquist, 2019). Reflecting this, many of studies have found a positive relationship between language and emotion development (Cole et al., 2010; Harris et al., 2005; Pons et al., 2003). This association appears consistent for preschool and school-aged children, spanning a range of aspects of both language and emotion understanding (Beck et al., 2012; Bosacki & Moore, 2004; Ornaghi & Grazzani, 2013; Pons et al., 2003; Streubel et al., 2020). For emotion recognition, evidence for associations with language stems predominantly from research involving facial expressions. Indeed, associations between general language ability and facial emotion recognition accuracy operate independent of age, and a range of other neurocognitive capacities such as theory of mind and attention/executive function (Beck et

al., 2012; Rosenqvist et al., 2014). These findings support an integral role for language in typical facial emotion recognition development.

Research examining associations with vocal emotion recognition is limited, as is research considering a wider range of participants beyond typical development. However, findings point to a longitudinal relationship between early language ability and both facial and vocal emotion recognition ability later in childhood (Griffiths et al., 2020). This effect was apparent over and above non-verbal IQ, in line with past research (Rosenqvist et al., 2014). Further, children with Developmental Language Disorder (DLD) – a language disorder not associated with a primary difficulty with socioemotional processing – also display difficulties with vocal emotion recognition (Griffiths et al., 2020), while targeted vocal emotion recognition intervention may have positive developmental implications for children with DLD (Durgungoz & St Clair, 2024). Collectively, these findings suggest a specific and unique relationship between language and emotion recognition development that extends to vocal stimuli. Similarly, for music stimuli, one study found that general verbal ability was correlated with musical emotion recognition accuracy in 3-6-year-old children (Franco et al., 2017). However, beyond this study, there is a lack of research exploring similar associations between language and musical emotion recognition development. Considering links to condition-general emotional processing mechanisms at acoustic-perceptual and cognitive levels (Escoffier et al., 2013; Lima et al., 2016; Vigl et al., 2024), it is plausible to hypothesise similar developmental mechanisms between audio conditions, too. Further, beyond these links, research points to a more fundamental developmental link between language and music.

1.6.2.2 Music and Language Across Development.

Research points to an inextricable link between the evolution of language and music, with overlap between the structural, cognitive, neurophysiological, and perceptual

components (Cross, 2014; Fedorenko et al., 2009; Jentschke, 2016; Patel, 2010). This is reflected in the ways in which musical and linguistic abilities develop. A systematic review explored interrelations between linguistic and musical abilities during early development (Pino et al., 2023). This review produced three key conclusions. First, early sensitivity to musical components of language, related to rhythm, is predictive of various components of language acquisition. Second, the ability to process and understand melody is predictive of language abilities such as the ability to recognise emotion in vocal prosody. Third, home musical environment – including regularity of infant directed song - predicts later language development. This suggests that links between language and musical emotion development may not only be a function of secondary links to vocal emotions, but also a primary developmental link between musical and linguistic development. These developmental language-music links also further signal the importance of considering singing stimuli within cross-condition developmental research. Indeed, comprehension of emotional components of singing may form the basis of later development of emotion understanding for both instrumental music and vocal prosody (Pino et al., 2023; Schubert & McPherson, 2015), while various aspects of the early home musical environment are causatively related to better language outcomes later in development (Franco et al., 2022; Papadimitriou et al., 2021; Politimou et al., 2019).

Collectively, findings suggest bidirectional interactions between various aspects of musical and linguistic development during childhood. Considered alongside domain-general socio-emotional processing mechanisms for music and voice (Escoffier et al., 2013; Frühholz et al., 2016), evidence indicates that musical emotion recognition development may be associated with language development in a similar fashion to facial and vocal domains (Franco et al., 2017). However, the mechanisms that underpin links between language and

emotion recognition are less clear. Recent research has suggested that the mediating role of developing *emotion-specific* language comprehension may be key in this regard.

1.6.2.3 Getting Specific – Moving Beyond General Language Abilities.

One mechanism through which language may relate developmentally to emotion recognition is via emotion concept development (Shablack & Lindquist, 2019). Language-mediated emotion concepts allow children to organise knowledge of different emotions, including their possible causes, consequences, forms of expression, and affective correlates (Barrett, 2006, 2017; Lindquist, 2017). Emotion language comprehension develops rapidly across childhood, with some evidence suggesting that TD children's emotion vocabularies almost double every two years between the ages of 4 and 11 (Baron-Cohen et al., 2010). Although the extent to which these emotion concepts constitute, or interact with, emotion perceptions remains disputed (Satpute & Lindquist, 2021), developing emotion language comprehension facilitates maturation in a range of emotional skills (Nenchevca et al., 2023). Indeed, children's emotion vocabulary predicts their general level emotion understanding independent of age, sex, and general verbal ability (Orghani & Grazzani, 2013).

Importantly, emotion-specific language development may facilitate a shift in emotion recognition from a broader dimensional understanding, based on arousal and valence, to the ability to categorise a wider range of emotions (Widen & Russell, 2008; Woodward et al., 2021). Supporting this assertion, research suggests positive correlations between children's emotion-specific vocabulary and facial emotion recognition accuracy (Beck et al., 2012), even when controlling for general language ability (Streubel et al., 2020). This suggests a specific association between emotion-specific language and emotion recognition development, in line with predictions that emotion language understanding facilitates more mature emotion perceptions via concept development (Shablack & Lindquist, 2019). However, due to a narrow focus on typical development within past research, it is currently

unclear whether links between emotion language and emotion recognition accuracy would extend to children with more varied levels of socioemotional difficulties.

There is limited research examining associations between emotion-specific language and audio emotion recognition development. However, one study found that instrumental recognition accuracy was predicted by emotion-specific verbal fluency in 5-6-year-old children, while no such relationship was apparent for general verbal fluency (Plate et al., 2022). This aligns with research on facial emotion recognition and suggests a possible domain-general role for emotion language comprehension in emotion recognition development. No similar research has examined the relationship between emotion-specific language and vocal emotion recognition. However, given the association between emotion-specific language and instrumental recognition accuracy (Plate et al., 2022), and the mediating role of emotion-specific language in links between general language ability and facial emotion recognition development (Streubel et al., 2020), an association with vocal emotion recognition is plausible. Indeed, emotion-specific language ability may operate as one of the proposed domain-general socioemotional mechanisms underpinning similar neurophysiological and perceptual patterns between musical and other forms of emotion stimuli (Escoffier et al., 2013; Lima et al., 2016; Paquette et al., 2018; Proverbio et al., 2020).

1.6.3 A Cross-Condition Model Through Which to Understand Audio Emotion Recognition Development

Figure 1.5 represents a cross-condition model of audio emotion recognition, synthesising the theoretical and empirical research discussed above. It is based on Bestelmeyer's (2014) multi-stage emotion recognition model, but also incorporates the theorised interactive role of fundamental arousal and valence dimensions, and emotion categories, during emotion recognition. Within each processing stage, possible condition-general and condition-specific (music-specific or voice-specific) aspects are considered.

Emotion recognition development can then be considered as age-related change within the different components and stages of the model, as well as in the interactions between them.

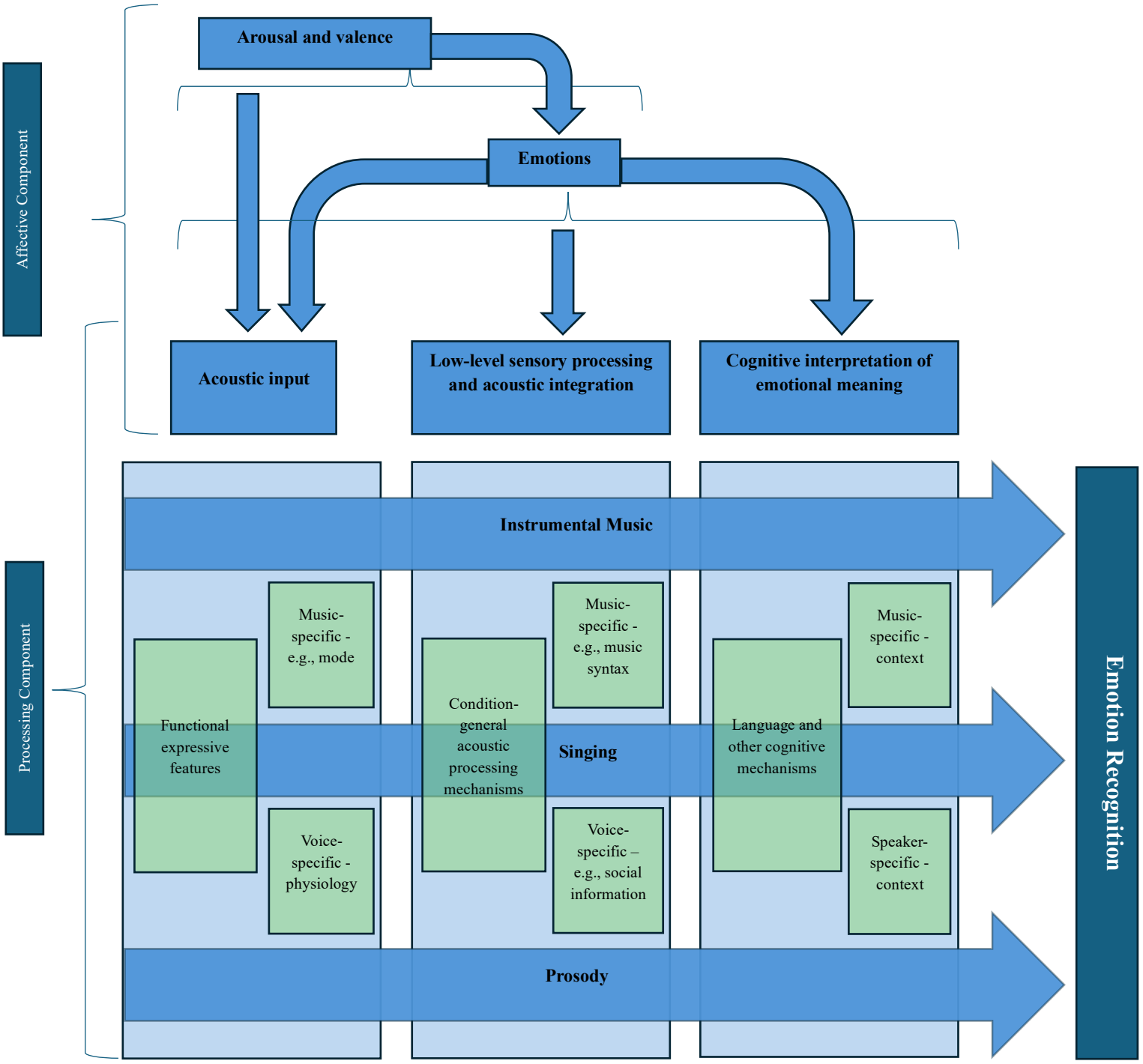
The model comprises the stages acoustic input, sensory processing/acoustic integration, and cognitive emotion interpretation. The ways in which these stages operate is organised within two components – the 'affective component' and the 'processing component'. The 'affective component' displays interactions and independent associations between dimensions arousal and valence, and emotion categories, during the different stages of the emotion recognition process. At the acoustic input stage, arousal and valence operate as low-level components of emotion categories. However, arousal and valence dimensions, and emotion categories, can also operate independently. This aligns with the integrated theoretical framework discussed (section 1.3.5), with evidence that both emotion categories and arousal properties show cross-cultural salience in audio emotion expressions (Athansopolous et al., 2021; Cowen et al., 2019), and that acoustic features can express emotions independent of arousal and valence (Bänziger et al., 2015). At the low-level perception/integration and cognitive interpretation stages, arousal and valence properties are integrated via their relationships to emotion categories. As noted, although these properties may have independent associations with certain emotion processes (e.g., expressed arousal acting as an attention-grabbing filter - Holz et al., 2021), at the perceptual level, emotion categorisation is key to decision-making (Markman & Rein, 2013) and to the development of various socio-emotional abilities (Chronaki et al., 2015a; Cooper et al., 2020). Developmentally, the proposed shift from recognising emotions based on affective dimensions arousal and valence to a more advanced understanding of emotion categories (Ruba et al., 2018; Widen & Russell, 2002; Widen 2008), will relate to a shift in the relative strength of relationships between these affective features and the cognitive interpretation of emotions (Woodward et al., 2022).

The ‘processing component’ outlines the ways in which these affective components are processed during emotion recognition, and possible condition-general and condition-specific aspects therein. At the acoustic input stage, this involves functional ‘iconic’ links between expressive features of music and voice, in line with Juslin’s (2013) model (Figure 1.1). Here, music-specific features, such as musical mode and melodic progressions, are pertinent to instrumental and singing stimuli, while voice-specific features relating to the physiological constraints of the voice are specific to prosody and singing stimuli. At the low-level processing/integration stage, individual differences in the sensitivity to acoustic properties (i.e., musical aptitude - Jansen et al., 2023; Vigl et al., 2024) such as those seen in individuals with congenital amusia (Lima et al., 2016), will affect processing in a condition-general fashion. Music-specific mechanisms at this stage may include processing of musical syntax (as in the ‘intrinsic’ layer of Figure 1.1) or the proposed music-specific pitch processing system (Zatorre & Baum, 2012). Voice-specific aspects in this stage relate to its tendency to refer to other social information, which may affect which emotionally significant acoustic information is integrated (Reybrouk & Podlipniak, 2019).

Finally, the cognitive emotion interpretation stage may involve condition-general mechanisms (Escoffier et al., 2013; Lima et al., 2016). As discussed, one such mechanism that may be particularly relevant developmentally is language, although it is currently unclear whether general language or emotion-specific language is likely to have the most direct role in audio emotion recognition (Griffiths et al., 2020; Streubel et al., 2020). Although this mechanism is presented as succeeding low level acoustic mechanisms, as noted, language may play a more fundamental role in the integration of emotionally meaningful acoustic information if assumed to partially constitute emotion perceptions (Satpute & Lindquist, 2021). Regardless, individual differences at this cognitive level are likely to have relatively strong influences on emotion recognition accuracy. However, context will continue to

modulate condition-specific inferences, such as those related to musical associations (Juslin, 2013), and speaker overall expressiveness (Woodward et al., 2021). Further, culture will modulate various aspects of the model, such as the nature of music-specific conventions at the acoustic input and integration stages (Athansopolous et al., 2021; Stewart & Walsh, 2005), and more condition-general culture-specific language effects at the cognitive interpretation stage (Lindquist, 2017).

Figure 1.5 – Cross-Condition Model of Audio Emotion Recognition



This model outlines how children's development within different stages may be important for their emotion recognition development. In doing so, it also outlines the levels at

which individual differences may arise. However, it does not outline possible implications of individual differences in audio emotion recognition in relation socio-emotional development, more broadly.

1.7 Emotion Recognition Difficulties and Broader Socioemotional Dimensions

1.7.1 A Transdiagnostic Approach to Individual Differences

Neurodevelopmental difficulties are often considered within diagnostic categories. Within this approach, diagnostic systems, most notably the Diagnostic and Statistical Manual of Mental Disorders (DSM-5 – American Psychiatric Association, 2013), are formulised and applied to categorise individuals' difficulties. Such diagnoses encompass a range of learning, communicational, and behavioural difficulties, including autism, attention-deficit hyperactivity disorder (ADHD), conduct disorder (CD). Importantly, these diagnostic categories predicate the nature and extent of support/interventions individuals receive (Astle et al., 2022).

While categories continue to evolve as new evidence emerges, this approach has received criticism. For example, it is claimed that diagnostic categories lack sensitivity to lower level, yet still functionally significant difficulties in many of the key underpinning socio-emotional and behavioural components (Astle et al., 2022). Further, it is claimed that categorical diagnoses can overlook the individualised nature of difficulties and associated support needs (Lahey et al., 2022). Indeed, individuals often experience difficulties that align with multiple diagnostic categories simultaneously (Lahey et al., 2021), while these difficulties often shift across development (Shevlin et al., 2017). However, this is not reflected in the research base, with most research focused on singular diagnoses (Astle et al., 2022). This may have detrimental implications, as children with more complex

neurodevelopmental profiles and/or those that fail to meet thresholds for a diagnosis are understudied and under-supported (Lahey et al., 2022).

This has given rise to an alternative, transdiagnostic approach to research into neurodevelopmental difficulties. This involves a dimensional approach, that considers the development of core mechanisms and characteristics that do not necessary align with specific diagnostic categories (Insel, 2014). Importantly, this dimensional approach to conceptualising psychopathology has strong empirical support (Haslam et al., 2020; Krueger et al., 2018). Further, hypothetically, it encourages developmental research that includes children with a wide range of difficulties, irrespective of their alignment with diagnostic thresholds. Practically, and pertinent to the present thesis, this means a difference in recruitment criteria. Specifically, recruitment is based on functionally defined needs – including children whose functional difficulties require informal or formal support at school (Landerl & Moll, 2010). This ensures that findings are more representative of the broader developmental difficulties experienced by the population of focus (Insel, 2014). The Neurodevelopment Assessment Unit (NDAU) at Cardiff University, through which some of the children within the present thesis were recruited, operates within this functional transdiagnostic framework.

1.7.2 Internalising Problems, Externalising Problems, and Child Outcomes

One of the most-adopted and well-supported transdiagnostic models of psychopathology is a 2-factor model including externalising and internalising problems (Eaton et al., 2015). Externalising behaviour refers to behaviours directed outwards to the social environment, including those relating to attention/hyperactivity, and aggressive, oppositional, and rule-breaking behaviours. Internalising behaviour encompasses behaviours with an internal focus, such as anxious and depressive symptoms, withdrawal, and somatic complaints (Achenbach & Edelbrock, 1978). Higher scores on each of these dimensions is related to poorer outcomes later in development. For example, children with higher

externalising problems are more likely to exhibit antisocial behaviour in adolescence (Bor et al., 2004) and have a poorer quality of life in adulthood (Korhonen et al., 2018). Similarly, children with higher internalising problems are more likely to self-harm and have higher levels of drug use in adolescence (Papchristou & Flouri, 2020), and experience depressive symptoms in adulthood (Korhonen et al., 2018). Within this framework, emotion recognition can be considered as a key underpinning ability that may relate to externalising or internalising presentations.

1.7.3 Vocal Emotion Recognition, Internalising, and Externalising Difficulties

Most research exploring associations between children's emotion recognition and these broad dimensions has involved facial expressions. While some research has found links with internalising-related difficulties (Easter et al., 2005; Schulz et al., 2001), others found no such link (Chronaki et al., 2015a) or one specific to certain emotions (Dede et al., 2021). For externalising, findings are more conclusive. A systematic review by Cooper et al. (2020) found evidence for emotion-general facial emotion accuracy difficulties in children with attention deficit hyperactivity disorder (ADHD), conduct problems, and callous-unemotional traits, particularly within samples of children with higher overall levels of externalising problems. However, when considered together, difficulties experienced by individuals with ADHD appear explainable by co-occurring CD, pointing to difficulties interpreting rather than attending to emotion stimuli as key to links between externalising problems and facial emotion recognition difficulties (Airdrie et al., 2018).

Despite extensive links between vocal emotion recognition and social competence in children (Morningstar et al., 2018; Nowicki & Mitchell, 1998; Verbeek, 1996), research considering links with internalising/externalising difficulties is limited. However, findings indicate some similarities with facial research. For example, research indicates a longitudinal association between vocal emotion recognition at age 8 and overall internalising and

externalising difficulties at age 10 (Nowicki et al., 2019). Further, in younger children with and without socioemotional difficulties, externalising, but not internalising, relates to both facial and vocal emotion recognition accuracy (Chronaki et al., 2015a). However, with TD samples, no link between vocal emotion recognition and either internalising or externalising problems has been found (Neves et al., 2021; Nowicki & Mitchell, 1998). It may be that, in line with research on facial expressions (Cooper et al., 2020), associations with externalising problems are stronger in groups of children with overall higher levels of externalising problems.

Some research has indicated that links between externalising problems and vocal emotion recognition are especially pronounced for attentional relative to behavioural aspects of the externalising dimension (Chronaki et al., 2015a; Sells et al., 2023). It is possible that attentional mechanisms play an increasingly integral role during vocal relative to facial emotion recognition, given the need to attend to and integrate complex acoustic information during emotion judgements (Bestelmeyer et al., 2014; Figure 1.5). However, there is a relative lack of vocal research accounting for co-occurring externalising-related conditions (Sells et al., 2023). Accordingly, it is also possible that more extensive research would highlight a more prominent role for difficulties interpreting, rather than attending to, emotion stimuli within links between externalising problems and vocal emotion recognition accuracy – in line with facial expression research (Airdrie et al., 2018). Irrespective of the mechanisms involved, evidence more strongly supports an association between vocal emotion recognition and externalising difficulties, relative to one with internalising difficulties. Further, given inconsistencies in the research base, research that maximises individual differences in socioemotional dimensions, while considering both broader internalising and externalising dimensions as well as sub-dimensions therein, may add clarity to the research base.

Current research in this domain does not extend to musical stimuli. As discussed, evidence points to extensive links between musical and vocal emotion expression and perception. Given further links between broader musical abilities and socioemotional development, a consideration of musical and vocal emotion recognition in tandem, in relation to socioemotional development, could be a fruitful avenue for research.

1.7.4 Music and Socio-Emotional Development

A growing research base links a range of musical abilities with various aspects of socioemotional development. In a recent systematic review, Blasco-Magraner et al. (2021) found a robust association between the educational use of music and children's emotion perception and regulation. These links were apparent for formal music training (e.g., Rose et al., 2015; Schellenberg et al., 2015) and in relation to specific musical interventions (e.g., Boucher et al., 2021). Importantly, musical interventions have been shown to reduce externalising-related behaviours such as physical and verbal aggression (Kim & Kim, 2018). Similarly, Gold et al.'s (2004) meta-analysis indicated that music therapy has a positive and substantial impact on socioemotional outcomes for children and adolescents - particularly for behavioural outcomes, relative to emotional ones. Isolating *emotional* components of musical ability, research also suggests a relationship between adolescents' ability to express emotions via music and their levels of self-reported externalising problems and empathy (Saarikallio et al., 2014). This, they suggested, indicates that participants' musical emotion understanding mirrors to some extent their broader abilities in socioemotional communication.

However, similar research focused on emotional components of music with younger children, or with those with socioemotional or behavioural difficulties, has not been conducted. Given music's powerful and ubiquitous ability to convey and elicit emotion (Day & Thompson, 2024; Eerola et al., 2013; Juslin & Sloboda, 2011), as well as its links to more general socioemotional processing mechanisms (Escoffier et al., 2013; Lima et al., 2016),

understanding of music at this level could be one mechanism through which these broader musical abilities relate to socioemotional dimensions. Accordingly, an exploration of how children with socioemotional difficulties' ability to perceive emotion in music relates to their broader socioemotional abilities could inform music-based interventions for this group.

1.7.5 Musical and Vocal Emotion Recognition in Relation to Externalising and Internalising Problems

When considered on emotional terms, possible common associations between music/voice and broader socioemotional dimensions emerge. Indeed, multiple reviews have indicated that music training is positively associated with vocal emotion recognition abilities (Martins et al., 2020; Nussbaum & Schweinberger, 2021). Importantly, an intervention focused specifically on emotion expression in music led to improvements in vocal emotion recognition in adults, even while controlling for formal music training (Mualem & Lavidor, 2015). This could point to emotional understanding of music via associated vocal emotion understanding as one avenue through which musical training/abilities relate to broader socioemotional dimensions.

However, as outlined in Figure 1.5, there are a range of possible emotion processing differences between conditions that may give rise to condition-specific associations with socioemotional dimensions. Indeed, the role of vocal prosody in directing attention to other social information may tie it more closely to these broader socioemotional constructs (Reybrouck & Podlipniak, 2019). For music, specific links have been identified between music processing and reward processing, while the importance of this reward system for positive socio-emotional adjustment is well established (Kasparek et al., 2020; Zatorre & Salimpoor, 2013). Further, children's home musical environment, and links to language development and positive developmental outcomes more generally, could have specific implications for links between musical emotion understanding and socioemotional

development (Politimou et al., 2018). For example, evidence suggests that musical engagement between caregiver and child supports positive attachment and emotional connection (Creighton et al., 2013; Fancourt & Perkins, 2018; Perisco et al., 2017). Caregiver-child attachment, particularly higher levels of insecure or avoidant attachment, is positively associated with externalising and internalising problems in children and adolescents (Achenbach et al., 2016). Accordingly, while musical and vocal emotion recognition may display similar links with broader socio-emotional dimensions, condition-specific links are also possible. Understanding these broader associations may be important in the context of potential music-based interventions and applications, particularly for children with socio-emotional difficulties.

1.8 Overview of Thesis, Research Questions, and Hypotheses

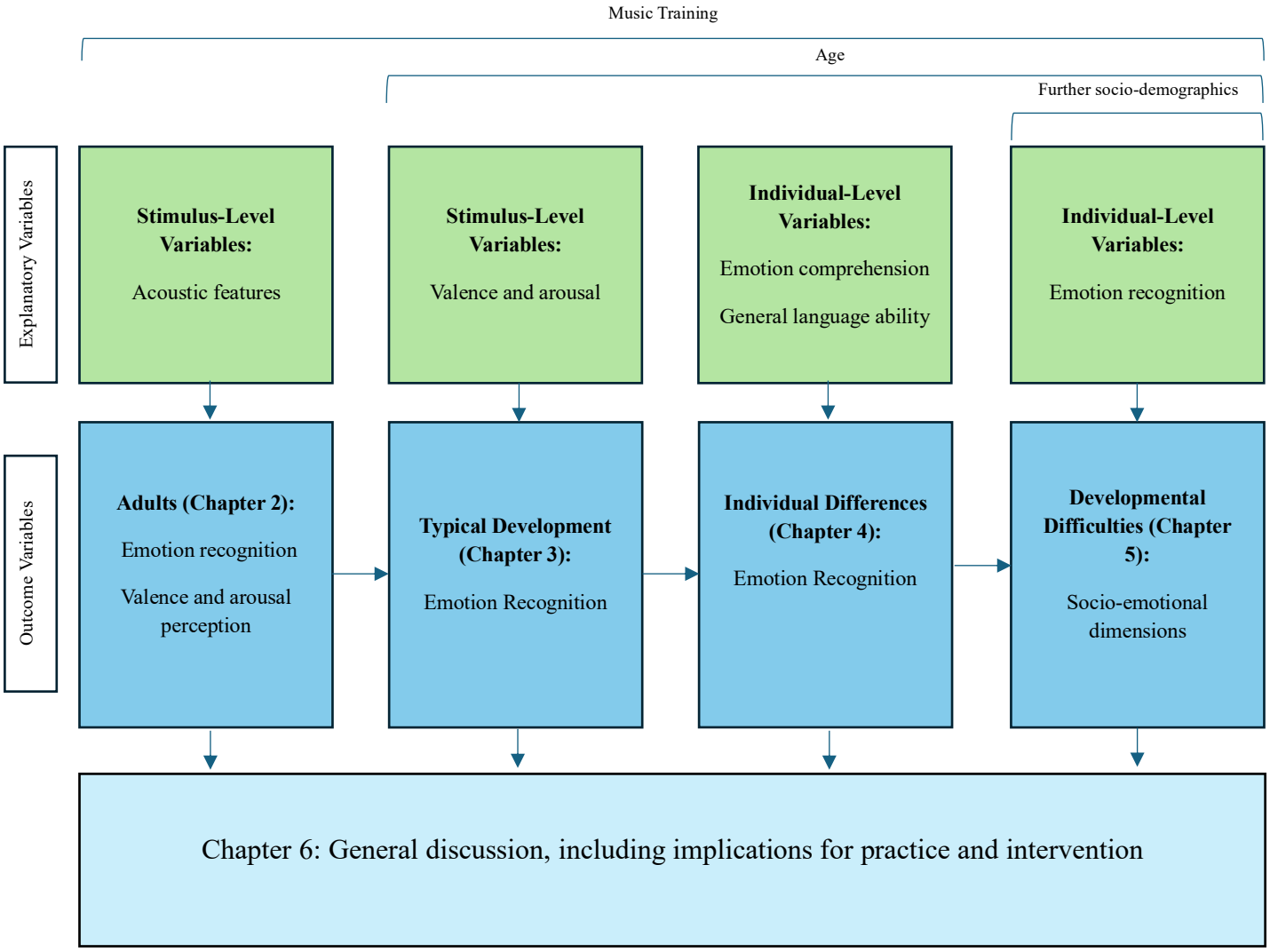
1.8.1 Thesis Overview

Despite extensive links between musical and vocal forms of emotion expression, little is known about the ways in which emotion recognition for these forms of audio stimuli develops, while research on individual differences in these processes is more limited still. Accordingly, the ultimate aim of this thesis was to contribute to a better explicated cross-condition model of audio emotion recognition through which to understand emotion recognition development, that accounts for differing forms of musical stimuli and considers individual differences in developmental processes. This could contribute to better evidenced musical applications/interventions centred on emotion recognition and socio-emotional outcomes more broadly.

Figure 1.6 outlines the structure of this thesis, with explanatory and outcome variables for each chapter. Chapter 2 focused on emotion perceptions, including discrete emotions and arousal/valence, in adults. This was the first study to include instrumental, singing and

prosody stimuli in tandem. Stimulus acoustic features were also considered in relation to perceptual patterns. The aim was to illuminate the nature of perceptual similarities and differences between musical and vocal conditions, in relation to the integrated theoretical model of vocal and musical emotion expression discussed (section 1.3.5). Chapter 3 shifted focus to emotion recognition in TD children. Associations between stimulus-level features at the affective level (valence, arousal) and emotion perceptions were also explored, considering the apparent importance of these fundamental features for emotion recognition development (Woodward et al., 2022; Cespedes-Guevara & Eerola, 2018). Chapter 4 focused on individual differences in children's audio emotion recognition. Language (general and emotion-specific) was considered as a possible condition-general mechanism of emotion recognition development, in line with research stressing the importance of emotion concept development for emotion recognition (Shablack & Lindquist, 2019). The sample included children with varied levels of socioemotional and behavioural difficulties, to ensure findings were applicable to children with a broad range of socioemotional functioning. Chapter 5 considered links between emotion recognition and broader socioemotional dimensions of externalising and internalising behaviours, in children with higher levels of socioemotional difficulties, before a discussion of findings and their wider implications in Chapter 6.

Figure 1.6 – Thesis Structure, Including Explanatory and Outcome Variables



1.8.2 Thesis Research Questions

1. Do adults' perceptions of emotions, and valence and arousal, differ between musical and vocal stimuli, and how do patterns relate to stimulus acoustic features?
2. In TD children, are emotion recognition patterns similar for musical and vocal stimuli, and how do patterns relate to stimulus-level arousal and valence dimensions?

3. Are there individual differences in children's emotion recognition accuracy for musical and vocal stimuli, and can these be partially explained by general and/or emotion-specific language development?
4. Is emotion recognition associated with broader socio-emotional dimensions in children with high levels of difficulties in these domains, and are these associations similar between musical and vocal conditions?

1.8.3 Key Hypotheses

The following broad hypotheses were made, but each chapter provides more detailed hypotheses that are specific to the analyses conducted. Given cross-over in research questions and analyses, some of these hypotheses are relevant to multiple chapters.

1. Across chapters, adults' and children's emotion recognition accuracy will be correlated for instrumental, singing, and prosody conditions. This hypothesis was based on theoretical shared reliance of each form of emotion expression on variations in similar acoustic features to express emotions (Juslin & Laukka, 2003), and past research indicating similar emotion recognition accuracy, and the development of this ability, for instrumental and prosodic stimuli (Heaton & Allgood, 2015; Laukka & Juslin, 2007; Vidas et al., 2018). Given shared processing mechanisms underpinning the perception of emotion across conditions (e.g., Escoffier et al., 2013; Frühholz et al., 2016), and proposals that emotion comprehension of singing may form the basis of later developing emotion recognition for both instrumental and prosodic conditions (Pino et al., 2023; Schubert & McPherson, 2015), this pattern was expected to extend to singing stimuli.

2. The closest similarities between audio conditions will relate to arousal, in terms of adults' arousal perceptions for each emotion and associations between acoustic features and adults' arousal perceptions in Chapter 2, and associations between stimulus arousal and children's emotion perceptions in Chapter 3. This prediction was based on evidence suggesting that both musical and vocal stimuli are more saliently understood in terms of arousal, relative to both valence and distinct emotion categories (Holz et al., 2021), perhaps due to closer links between arousal and expressive acoustic features (Bänziger, et al., 2015).
3. Emotion-specific language will predict individual differences in children's emotion recognition accuracy across audio conditions in Chapter 4. Emotion-specific language has been shown to predict children's emotion recognition accuracy for music (Plate et al., 2022). This aligns with research on facial expression recognition (e.g., Streubel et al., 2020) and suggests a domain-general role for emotion language comprehension in emotion perception. Given shared socio-emotional mechanisms across musical and vocal stimuli (Escoffier et al., 2013; Lima et al., 2016; Paquette et al., 2018), and analogous links between general language and emotion recognition across modalities (Franco et al., 2017; Griffiths et al., 2020), similar associations were expected across audio conditions.
4. Children's musical and vocal emotion recognition accuracy will each show a negative association with the dimension of externalising difficulties in Chapter 5. Past research has indicated a relationship between vocal emotion recognition and socio-emotional adjustment, including externalising difficulties (Chronaki et al., 2015a). Concurrently, a range of musical abilities are related to positive socio-emotional developmental outcomes, including externalising behaviours (Blasco-Magraner et al., 2021; Kim & Kim, 2018). Accordingly, and given the shared processing mechanisms thought to

underpin emotion recognition for musical and vocal stimuli (e.g., Escoffier et al., 2013), common relationships were expected between emotion recognition for each condition and the externalising dimension.

Further analyses, including emotion recognition accuracy patterns for specific emotions (Chapters 2 and 3) and associations between acoustic features and participant's perceptions (Chapter 2), were analysed exploratively.

2. Adults' Perceptions of Affective Dimensions and Emotion Recognition in Instrumental Music, Singing and Vocal Prosody

2.1 Introduction

Music's ability to express and elicit emotions is well-established (Juslin, 2018; Juslin & Sloboda, 2011). Some have attributed music's emotional resonance to evolutionary links to vocal emotions, particularly vocal prosody – the suprasegmental aspects of speech including intonation, stress, and rhythm (Brown, 2017; Juslin & Sloboda, 2011). Indeed, overlapping neurophysiological correlates and cognitive processes between music and vocal prosody have been demonstrated (Escoffier et al., 2013; Frühholz et al., 2016; Paquette et al., 2018; Proverbio & Piotti 2022). Further, musical and prosodic stimuli each rely on variations in acoustic features such as loudness, pitch, and tempo/speech rate, to express emotions (Eerola et al., 2013; Grandjean et al., 2006; Juslin & Laukka, 2003; Scherer et al., 2015). However, theories disagree regarding the extent to which these expressive features align to specific emotions in a consistent manner between audio conditions. This has implications for how cross-condition emotion recognition research is approached, and the findings it produces.

2.1.1 Theories of Musical and Vocal Emotion Expression

Some theories claim that cross-condition similarities in perceptual patterns are due to close links between how specific musical and vocal emotions are expressed. For example, one prominent approach conceptualises music as a 'super-expressive voice' – an acoustically exaggerated vocal counterpart especially rich in emotional information (Juslin & Sloboda, 2011; Scherer, 1995). Links between music and voice within this perspective are tied to innate expressive features unique to each emotion (Juslin, 2013), in line with discrete emotion theories that claim that each emotion encompasses distinct physiological, subjective, and expressive components (Keltner et al., 2019). Meta-analytic evidence provides some

support for this theory, with similarities between musical and vocal emotions in terms of the acoustic cues communicating specific emotions (Juslin & Laukka, 2003). Other theories dispute the mapping of acoustic features to emotions in a consistent manner across audio conditions, instead suggesting that between-condition similarities relate to the expression of dimensions arousal and valence (Cespedes-Guevara & Eerola, 2018). Within this dimensional view, emotion recognition involves the interpretation of variations in these fundamental affective dimensions, and categorisation of the expressed emotion based on conceptual knowledge of emotion categories (Barrett, 2017; Bliss-Moreau & Barrett, 2009; Russell, 2003). For music, understanding of conventions such as musical mode also facilitates accurate emotion recognition (Dalla Bella et al., 2001). This perspective would predict more variation in emotion recognition patterns between music and voice, as expressive features do not align directly with emotion categories in a condition-general manner.

2.1.2 Similarities and Differences in Musical and Vocal Emotion Recognition Patterns – Links to Arousal/Valence Dimensions and Acoustic Features

Vocal emotion recognition involves multiple stages, including the decoding/integration of acoustic information, and the cognitive interpretation of emotional meaning (Bestelmeyer et al., 2014; Schirmer & Kotz, 2006). Each stage also appears relevant for musical emotion recognition, with similar neural correlates for musical and vocal stimuli demonstrated during acoustic integration (Peretz et al., 2015) and cognitive interpretation (Schirmer et al., 2012) of emotion stimuli.

Past audio emotion recognition research has provided mixed findings regarding the prominence of affective dimensions versus discrete emotions in musical and vocal emotion expressions. For example, there are some similarities between music and vocal prosody in adults' emotion recognition accuracy for specific emotions (Laukka et al., 2013a; Laukka et

al., 2013b), which appear consistent irrespective of levels of musical training (Correia et al., 2020; Vidas et al., 2018). Further, musical and vocal emotion recognition accuracy are significantly positively correlated in adults (Laukka & Juslin, 2007) and TD children (Heaton & Allgood, 2015; Vidas et al., 2018), while music processing difficulties are associated with difficulties for both musical and vocal emotion recognition (Lima et al., 2016). Collectively, these findings may indicate a fundamental link between music and voice based on expressive similarities.

However, past research has also highlighted condition-specific patterns of emotion recognition accuracy for some emotions. For example, anger is recognised accurately by adults and children in high intensity prosodic stimuli (Elfenbein et al., 2021; Sauter et al., 2013), but often difficult to recognise in music (Micallef Grimaud & Eerola, 2022; Vidas et al., 2018). Conversely, while happiness is recognisable in music from an early age (Dalla Bella et al., 2001; Vidas et al., 2018), difficulties recognising prosodic happiness can persist through adolescence (Chronaki et al., 2018) and into adulthood (Lausen & Hammerschmidt, 2020). This may argue against proposed alignment of expressive acoustic features to specific emotions in a condition-general fashion between music and voice, as proposed within discrete theories (Juslin & Laukka, 2003). From a dimensional perspective, some variation in emotion recognition patterns is expected, given that acoustic features align with arousal and valence rather than specific emotion categories (Cespedes-Guevara et al., 2018). It may be that arousal is particularly important to cross-condition similarities in emotion expression and perception. Indeed, past research suggests that arousal is more salient in musical and vocal stimuli relative to valence, both in terms of how easy it is to perceive and the condition-general acoustic features that relate to it (e.g., loudness and tempo/speech rate - Coutinho & Dibben, 2014; Holz et al., 2021; Llie & Thompson, 2006; Sauter et al., 2010).

A consideration of affective dimensions can also highlight expressive differences between audio conditions which may have implications for emotion recognition patterns. For example, major musical mode is robustly associated with positive valence, and minor mode with negative valence (Dalla Bella et al., 2001; Gomez & Danuser, 2007), which may make valence more salient in musical stimuli. This is also reflected in acoustic feature patterns, with acoustic features more strongly predictive of adults' perceived valence in instrumental music relative to vocal prosody (Weninger et al., 2013). This could contribute to certain differences in adults' emotion recognition patterns between conditions, such as the relative salience of happiness in instrumental stimuli relative to prosody (Micallef Grimaud & Eerola, 2021; Lausen & Hammerschmidt, 2020). This highlights the importance of considering affective dimensions, to explain both the similarities and differences in emotion recognition between audio conditions.

A final possible discrepancy between audio conditions relates to the consistency with which these affective dimensions substantiate emotion recognition patterns. Past research has indicated that perceptions of arousal and valence strongly predict emotion recognition for musical stimuli and vocal stimuli (Eerola & Vuoskoski, 2011; Sauter et al., 2010). However, there are also some acoustic properties that can communicate certain emotions, such as anger, independent of affective dimensions arousal and valence (Bänziger et al., 2015; Giordano et al., 2021). This could contribute to the higher recognition accuracy seen for prosodic anger relative to instrumental anger in past research (Sauter et al., 2013; Vidas et al., 2018). Accordingly, while consideration of valence and arousal dimensions may help to partially explain observed emotion recognition patterns, certain patterns may be explainable via more direct associations between expressive acoustic features and specific emotions. Consideration of a specific form of musical stimuli – singing – may be illuminating regarding the extent to which music and voice align based on affective dimensions and emotion categories.

2.1.3 Singing as a Conceptual Bridge Between Instrumental and Prosody Stimuli

Most past music emotion recognition research has focused on instrumental stimuli. For research considering cross-condition associations with vocal emotions, inclusion of singing stimuli may strengthen inferences. Indeed, there are commonalities between singing and vocal prosody in terms of adults' emotion recognition patterns and the acoustic features underpinning them (Scherer et al., 2015; Scherer et al., 2017). Acoustic features related to arousal appear particularly strongly related, possibly due to the physiological constraints pertinent to the speaking and singing voice (Scherer et al., 2017). Despite these links, emotions in prosody are generally easier than those in singing for adults to recognise (Livingstone & Russo, 2018). Singing is also uniquely tied to instrumental music; in that it is subject to musical conventions such as musical mode. Understanding of musical mode is important for emotion recognition (Eerola et al., 2013; Juslin & Laukka, 2003), and as noted, robustly related to perceived valence (Gomez & Daunser, 2007). Accordingly, given its physiological and arousal-based links to vocal prosody, and valence-based links to instrumental music, perceptual patterns for singing could be revealing regarding the nature of perceptual links between musical and vocal stimuli.

2.1.4 The Current Study

Despite the apparent importance of both arousal/valence dimensions and emotion categories to associations between musical and vocal emotions, past cross-condition research has not considered dimensional and discrete emotion perceptions in tandem. Further, past audio emotion recognition research has rarely included singing stimuli, and no studies have considered prosody, instrumental, and singing stimuli together. Given its unique relationship with aspects of both prosodic and instrumental emotion expressions, inclusion of singing stimuli could elucidate the nature of perceptual similarities and differences between musical and vocal stimuli. Incorporation of both discrete and dimensional perceptual patterns, and of

singing stimuli, may provide unique theoretical insights regarding the level at which audio conditions are most strongly related. Accordingly, the present chapter examined adults' perceptions of valence and arousal, and emotions, for instrumental, singing, and vocal prosody stimuli. A set of theoretically cross-condition acoustic features were also extracted and analysed in relation to perceptual patterns, to aid interpretation of similarities and differences between audio conditions.

2.1.4.1 Hypotheses.

1. Overall emotion recognition accuracy will be similar between instrumental and prosody audio conditions, with each higher than singing. Past research with adults has shown similarities in overall recognition accuracy for instrumental and prosody stimuli (Laukka & Juslin, 2007; Vidas et al., 2018), while prosody stimuli is easier to recognise than singing stimuli (Livingstone & Russo, 2018).
2. Accuracy for all audio conditions will be positively correlated. Past research points to shared processing mechanisms for emotion recognition across musical and vocal stimuli (Escoffier et al., 2013; Lima et al., 2016; Paquette et al., 2018). These overlaps suggest that individuals who are more skilled in decoding emotion in one condition are likely to be skilled in others.
3. Arousal perceptions for emotions will not differ between audio conditions. Arousal is generally perceived more easily than valence across musical and vocal stimuli, due to its strong association with universal acoustic features such as loudness, tempo, and speech rate (Coutinho & Dibben, 2014; Ilie & Thompson, 2006; Holz et al., 2021; Sauter et al., 2010). Because these cues are shared across all three audio conditions, arousal perceptions are expected to remain relatively consistent.

4. Valence perceptions will also be similar between conditions, but instrumental and singing stimuli will display more separation between negative and positive valence emotions, compared to prosody stimuli. There is evidence that music-specific expressive features, such as musical mode, enable greater separation of positive and negative valence emotions in musical stimuli (Gomez & Danuser, 2007; Dalla-Bella et al., 2001). In contrast, prosody stimuli may display more between-valence confusion due to subtler valence cues and the constraints of speech-based expression. This suggests that although mean valence ratings may be similar across conditions, instrumental and singing stimuli are more likely to support distinct perception of valence extremes.

Accuracy patterns for specific emotions were analysed exploratively, as were relationships with acoustic features. However, acoustic features were expected to be more consistently related to arousal across conditions, and these arousal-related features were expected to be most prominent within associations with emotion perceptions.

2.2 Methods

2.2.1 Participants

Participants were psychology undergraduate students from Cardiff University, recruited via the Experiment Management System. Of 223 participants that signed up, 159 were included in data analyses after data screening (see section 2.2.2.5 below), aged between 18 and 23 years (M = 19.5). Of the included participants, 133 identified as female, one as demiboy, one as transmasculine, and 24 preferred not to say. Participants were not deaf or hard of hearing and spoke English fluently. The study was granted ethical approval by the Cardiff University School of Psychology Research Ethics Committee and participants were awarded course credit.

2.2.2 Materials and Procedure

Participants completed the study online in a quiet space. Participants used headphones to provide control of sound presentation and attenuate environmental sounds (Woods et al., 2017). The tasks and questionnaire were created and presented via Gorilla experiment software (Anwyl-Irvine et al., 2020). Participants were first introduced to the study, including some general information regarding its aims and the tasks they would be completing. They then completed a task to ensure they were using headphones (details in section 2.2.2.5 below), followed by the emotion recognition task, the arousal and valence perception task, and the questionnaire. Participants could take breaks between tasks if desired. After completing all the tasks, participants were thanked and provided with more information about the study's aims, hypotheses, and possible implications. The whole study lasted approximately 45 minutes.

2.2.2.1 Emotion Recognition Task.

2.2.2.1.1 Materials.

Prosody and singing stimuli were a randomly selected set of four female speakers from the validated Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) stimuli set (Livingstone & Russo, 2018). Each speaker expressed emotions anger, fear, sadness, calmness, and happiness at high intensity, in both spoken and sung forms. This provided 20 stimuli for prosody and singing conditions. Each stimulus type involved the same semantically neutral sentence – 'dogs are sitting by the door'. Singing stimuli were performed in either major (happiness, calmness) or minor (sadness, fear, anger) musical modes. Stimuli were between 3-5 seconds in length. Female only speakers allowed focus on between-emotion differences without introducing further variation in emotionally relevant acoustic features (e.g., pitch) that are generally greater between sexes (Kamiloglu et

al., 2020; Nussbaum et al., 2021). High intensity stimuli were chosen to match instrumental music – composed and performed to maximally communicate the given emotions via a single instrumental timbre (piano).

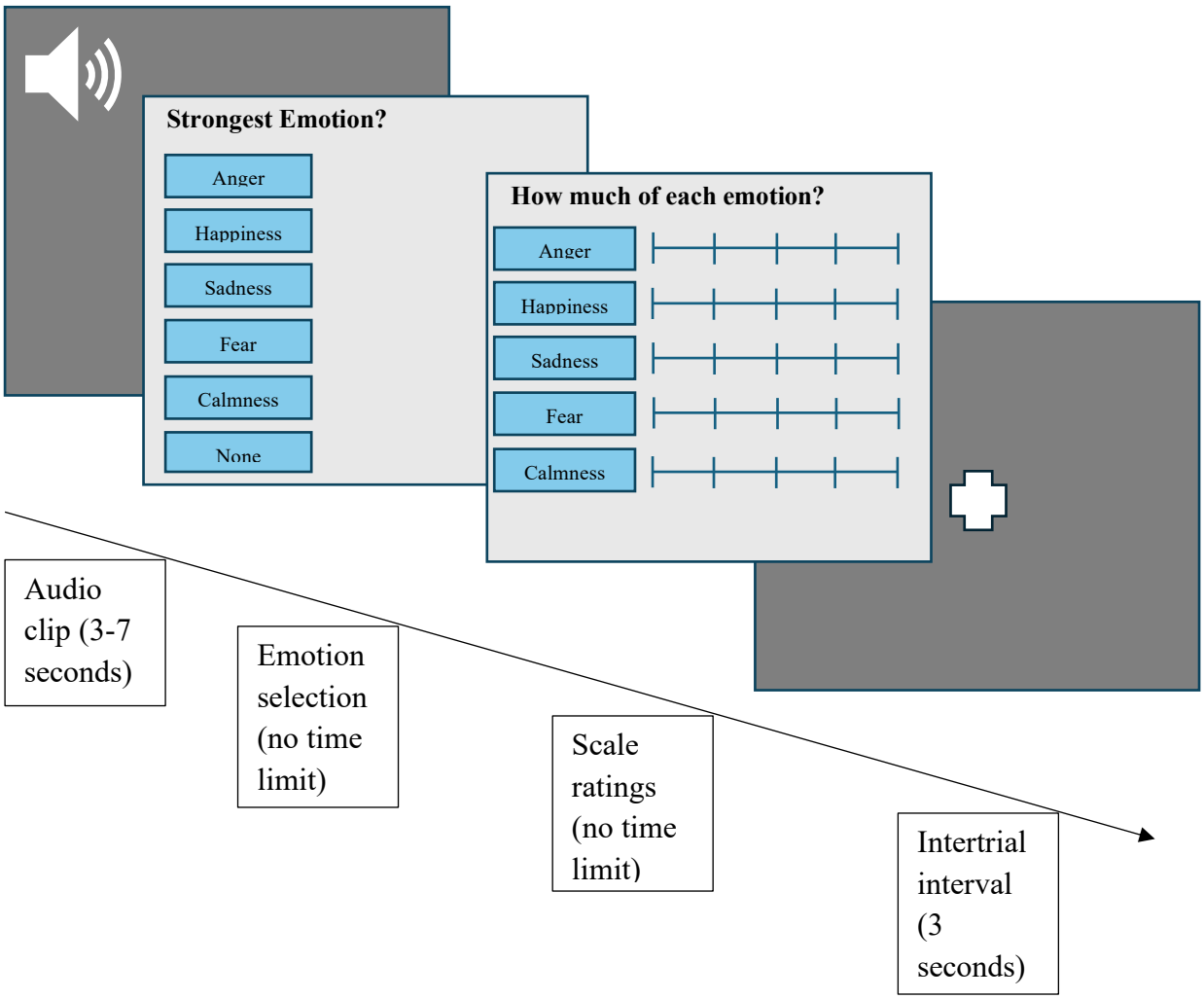
Instrumental stimuli were 20 piano pieces expressing the same emotions as above. Three pieces for each emotion were taken from Micallef Grimaud & Eerola's (2021) validated set. These stimuli balance ecological validity and novelty via original composition. Further pieces expressing happiness, sadness, fear, and calmness were taken from Vieillard et al.'s (2008) validated set of computer-generated piano pieces. The final anger piece was taken from Sutcliffe et al.'s (2017) validated set. Stimuli from these additional sets were re-recorded by a professional musician, using the same recording parameters adopted for Micallef Grimaud & Eerola's (2021) set. Stimuli were between 4-7 seconds in length, in keeping with the 'mental presence time' for audio stimuli and to match vocal stimuli as closely as possible (Argstatter, 2016). Gating paradigms have indicated that the lower bound of this range is sufficient to achieve maximal emotion recognition accuracy for musical excerpts (Vieillard et al., 2008).

Stimulus loudness was normalised as far as possible within and between conditions (see Appendix A for details). Participants' own volume levels were normalised by presenting a looping piano scale increasing in equal steps from the minimum to 0.2 dBFs above the maximum sound level within the stimuli set. Participants were asked to set their sound to zero and increase the level until it was as loud as possible without becoming uncomfortable, to allow optimum perception of variations in loudness across the stimuli set. This was important, as loudness is a condition-general communicator of arousal (Llie & Thompson, 2006; Weninger et al., 2013).

2.2.2.1.2 Procedure.

For the emotion recognition task, each trial began with an automatically triggered audio clip. Participants then made an emotion categorisation (which emotion was expressed 'most strongly' in the stimulus), followed by scale ratings of how much of each emotion they perceived (Figure 2.1). This allowed for the possibility of ambiguity in participants emotion perceptions. There was no time limit for trials, and trial length varied based on stimulus length and how long participants took to make their judgements. There was a three second interval between trials during which the screen displayed a white cross on a dark grey screen, before the next audio clip was automatically triggered.

Figure 2.1 – Schema for Emotion Perception Task



Audio conditions were presented separately and in a random order, and stimulus presentation order was randomised within each condition. For each participant, accuracy scores and scale ratings, per emotion and condition, were produced.

2.2.2.2 Arousal and Valence Perceptions.

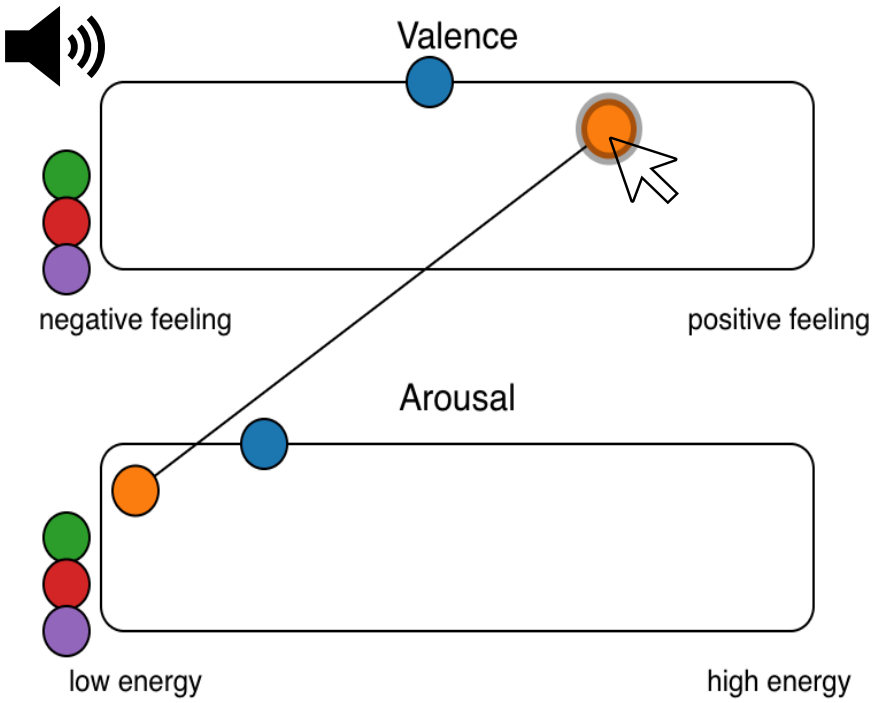
2.2.2.2.1 Materials.

The stimuli for the arousal and valence perceptions task were the same as those adopted for the emotion recognition task.

2.2.2.2.2 Procedure.

In the arousal and valence perception task, participants rated the perceived valence (negative to positive feeling) and arousal (low to high energy) for each stimulus. As this task did not involve an accuracy judgement against a pre-defined category, the task adopted the format of the Audio-Tokens Toolbox (Figure 2.2 - Donhauser & Klein, 2022). Participants hovered over the tokens to hear the stimulus and dragged them along each scale to rate their level of valence and arousal.

Figure 2.2 – Schema for Valence and Arousal Perception Task



This approach overcame common issues in sequential audio perception studies relating to participants 'drifting' in their ratings based on earlier presented stimuli (Gerratt et al., 1993). It also reduced demands on participants, including those on working memory, with stimuli playable multiple times and positioned relative to other stimuli. There were 12 trials - four for each audio condition. Within each trial, there were five stimuli for participants to listen to and rate on valence and arousal in relation to the other stimuli presented. There was no time limit set for trials, and there was a five-second interval between trials, within which a white cross was displayed on a dark grey screen. Conditions and the trials therein were fully randomised and counterbalanced, as was stimulus position within each trial.

2.2.2.3 Questionnaire.

Participants completed a questionnaire asking for demographic information and their years of formal instrumental and music theory training. Music training questions were

combined to form a single 'music training' variable. Details of the structure of this music training variable are in Appendix G.

2.2.2.4 Stimuli Acoustic Feature Levels.

Acoustic features were extracted via openSMILE software (Eyben et al., 2010). Features were selected based on theoretical significance for both musical and vocal stimuli and past cross-condition research (e.g., Eerola et al., 2013; Gabrielsson & Lindstrom, 2010; Juslin & Laukka, 2003; Llie & Thompson, 2006; Paquetta et al., 2018). Features were then correlated to ensure they were distinct and to avoid issues of multicollinearity in later analyses. This produced a set of six distinct features – mean loudness, loudness variation, mean pitch, pitch variation, tempo/speech rate, and brightness. Coefficients of variation, rather than standard deviation, were employed as measures of variability due to standard deviation's correlation with mean levels. The acoustic features were able to predict the emotion category of stimuli significantly above chance for all audio conditions (see Appendix B). More detail on acoustic feature selection, extraction, and raw feature levels for each stimulus within each condition are available in Appendix B. Z-scored feature levels, adopted for certain analyses, are available in Appendix C.

2.2.2.5 Data Screening.

Data were screened to ensure a controlled audio environment, and against careless responding - related to higher error variance (Maniaci & Rogge, 2014). This involved four steps:

1. Design factor – headphone check task (Milne et al., 2021). Participants heard three white noise sounds and had to choose the correct option that contained a tone, which was only passable with headphones. Those that scored below chance were removed.

2. Administration factor – speed of completion (excluding breaks between tasks). Outliers on the low end of the distribution that completed the tasks very quickly were removed.
3. Statistical factor 1 – providing the same response consistently in the emotion recognition task ('longstring' approach - Meade and Craig, 2012). Outliers at the top end of the distribution for both maximum and average number of consistent responses, across at least two audio conditions, were removed.
4. Statistical factor 2 – providing random responses in the valence and arousal task (psychometric synonyms - Meade & Craig, 2012). Responses for individual stimuli were correlated across the whole sample, and participants that diverged significantly from this sample-level pattern for both arousal and valence ratings were removed.

2.2.3 Statistical Analysis

In line with recent suggestions for psychological research (Keysers et al., 2020; Quintana & Williams, 2018), Frequentist analyses were conducted, and where relevant and informative, Bayesian results were also included in the form of Bayes Factors (BFs). BFs work via direct comparison of two competing models for the data – one with the effect of interest and one without (Morey et al., 2016). BFs have the advantage of providing direct evidence for both alternative and null hypotheses – important for the present research which aims to quantify differences *and* similarities between different audio conditions. They were interpreted in line with Jeffrey's (1998) specifications, outlined in Appendix D. Inferences were based on coherence between *p*-values and BFs (e.g., evidence against an effect included a *p*-value >.05 and a BF <1/3). Where Bayes Factors and Frequentist statistics led to different conclusions, results were interpreted primarily using Frequentist statistics (p-values), given their broader availability across all analyses. All analyses were conducted in R (R Core Team,

2022). Full details of statistical procedures, including assumption tests, specific R packages, and reporting, can be seen in Appendix E.

2.2.3.1 Frequentist Analyses.

2.2.3.1.1 Emotion Recognition.

A Generalised Linear Mixed Model (GLMM) with correct recognition as a binary dependent variable (1/0) was conducted to analyse emotion recognition patterns. Independent fixed effects were condition, emotion, and the interaction between condition and emotion. Participant and stimuli were entered as random intercepts (see Appendix F for model selection procedure). Due to the binary dependent variable, post-hoc comparisons were analysed in terms of odds ratios. A correlation analysis was also conducted between mean emotion recognition accuracy for each audio condition, and between each condition and music training.

2.2.3.1.2 Valence and Arousal Perceptions.

Linear Mixed Models (LMMs) were employed to explore differences in arousal and valence ratings (rated on a continuous scale from 0-100) by emotion and condition. For the valence model, whether participants had any music training (entered as a categorical yes/no variable) was also included. Participant and stimuli were included as random intercepts. For valence, a by-participant random slope was included for emotion, and for arousal, slopes were included for emotion and condition (see Appendix F for model selection process).

2.2.3.1.3 Acoustic Feature Patterns.

Acoustic features were analysed in relation to perceptual patterns. For valence and arousal, correlation analyses were conducted between each stimulus' average valence and arousal rating, and raw acoustic feature levels for each acoustic feature, within each audio condition. A further LMM was conducted for both valence and arousal, to compute the total

variance in arousal and valence ratings (R-squared value) explained by the whole set of acoustic features, while accounting for random variance at the stimulus and participant levels. For emotion perception data, the proportion of time each emotion was selected was correlated with z-scored acoustic feature levels for each stimulus, within each audio condition. Standardised z-scores were used in this case to account for how each speaker differentiates emotions. Spearman's coefficients were adopted for all acoustic analyses. A series of GLMMs with emotion selection (yes/no) as a binary dependent variable and acoustic features as independent variables were conducted to produce marginal R-squared values for each emotion, within each condition. This outlined the amount of variance in emotion selections explained by the whole set of acoustic features, while accounting for random variance at the participant and stimulus levels.

2.2.3.2 Bayesian Analyses.

BFs were included for a) correlation analyses, b) LMM/GLMM analyses at the level of fixed effects (including post-hoc comparisons). All Bayesian models were fit with matching fixed and random effects structures to the frequentist models above, using the default priors established in past research (Morey et al., 2015; Oberaur, 2019, 2023). For mixed models, they were not included for random effects as these were not specified with default priors, as is required for accurate bayes factors. Further details of Bayesian analyses, including prior specifications, can be seen in Appendix E.

2.3 Results

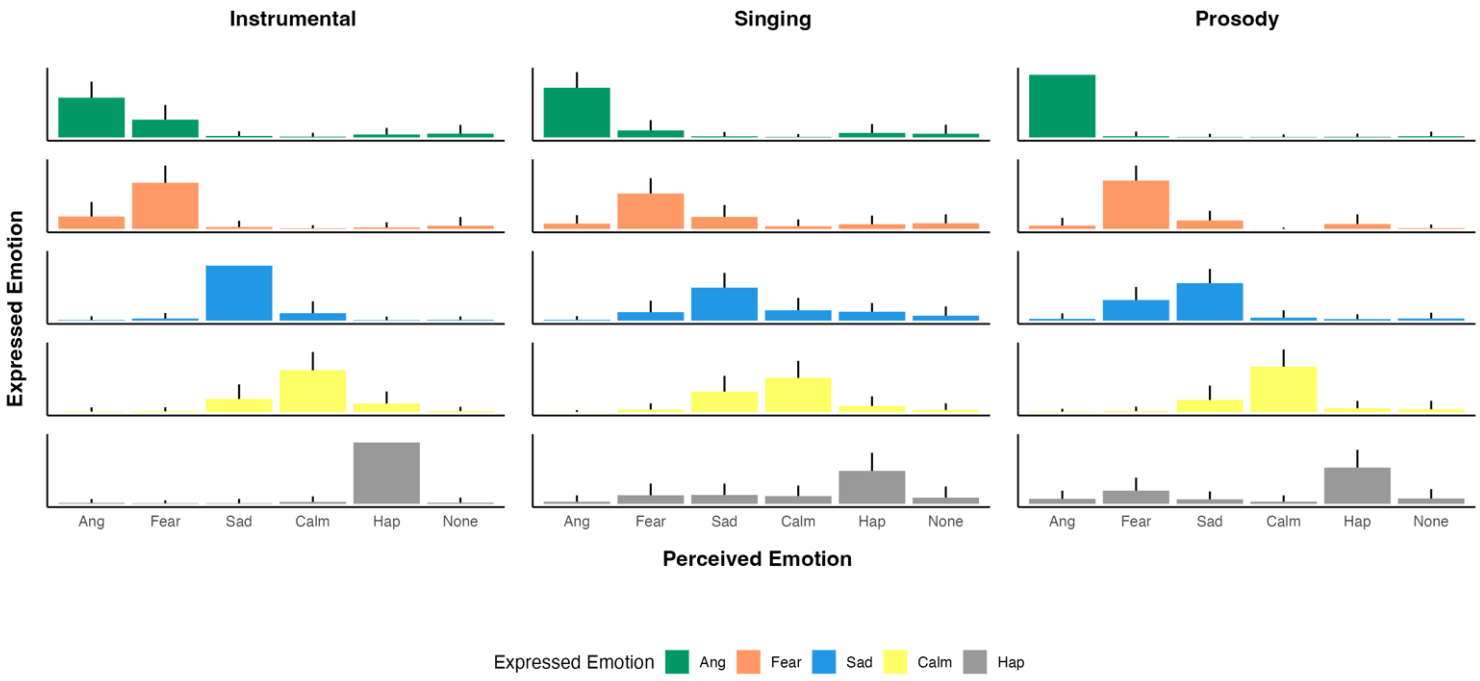
2.3.1 Emotion Recognition Accuracy by Condition and Emotion

Figure 2.3 presents emotion recognition accuracy (%) and patterns of confusion for each audio condition[8]. Most prominent confusion patterns were between emotions

[8] Raw emotion recognition scores and confusion patterns in table form are in Appendix H.

hypothetically similar in arousal. Calmness was consistently categorised as sadness, there was a bidirectional confusion between anger and fear for instrumental music and happy prosody was confused with fear. However, sad prosody was also confused with fear, and singing stimuli was confused more widely. Scale ratings of emotions were also collected, but patterns were almost identical to emotion recognition patterns, suggesting that emotions were generally perceived unambiguously. These can be seen in Appendix J.

Figure 2.3 - Raw Mean Emotion Recognition Accuracy (%) and Confusion Patterns by condition and emotion.



Note. Error bars = standard deviation. The main diagonal represents correct emotion recognition. Ang = anger; Sad = sadness; Hap = happiness. *N* = 159. X-axis = % correct.

A GLMM was used to analysed emotion recognition patterns, with 'correct recognition' (0/1) as the dependent variable and emotion and condition, and their interaction, as independent variables. Stimuli and participant were entered as random intercepts. The

random intercept for participant significantly improved model fit (c^2 (1) = 80.252, p<.001), as did the intercept for stimuli (c^2 (1) = 811.18, p<.001).

There was a significant main effect for condition (c^2 (2) = 12.85, p =.002, BF_{10} = 3.49). The odds of a correct response for instrumental music were significantly higher than for singing (OR = 2.70, *SE* = 0.67, p<.001, BF_{10} = 15.15), and prosody had higher odds than singing (OR = 2.38, *SE* = 0.59, p = .001, BF_{10} = 5.79). The odds for instrumental music and prosody did not differ (p =.45, BF_{10} = 0.21).

There was also a significant interaction between condition and emotion (c^2 (8) = 24.45, p =.002, BF_{10} = 29.14). The odds of a correct response for happiness were greater for instrumental music than prosody (OR = 11.59, *SE* = 6.55, p =.001, BF_{10} = 7.48) or singing (OR = 14.32, *SE* = 8.09, p<.001, BF_{10} = 30.72). Sadness also displayed markedly higher odds of a correct response for instrumental music compared to prosody (OR = 3.47, *SE* = 1.93, BF_{10} = 0.41) and singing (OR = 6.03, *SE* = 3.35, BF_{10} = 2.64), but these did not reach statistical significance (*ps*>.05). Anger had higher odds of a correct response for prosody compared to instrumental music (OR = 13.04. *SE* = 7.43, p<.001, BF_{10} = 61.15). Odds were also higher than singing, but while the BF strongly supported this difference, it did not reach statistical significance (OR = 6.51, *SE* = 3.73, p =.07, BF_{10} = 25.28). Accuracy for fear and calmness did not differ between conditions (ps >.05, BF_{10s} < 1).

2.3.2 Associations Between Emotion Recognition Accuracy and Music Training

Correlations between emotion recognition accuracy for instrumental, singing, and prosody conditions, and years of music training, are in Table 2.1. Music training was not associated with recognition accuracy for any audio condition, with anecdotal evidence against a correlation with instrumental music, and moderate evidence for a lack of correlation for

both singing and prosody[9]. Emotion recognition accuracy was positively correlated between all audio conditions.

Table 2.1 – Correlations Between Music Training and Recognition Accuracy

Variable	1	2	3
1. Accuracy Instrumental	-		
2. Accuracy Singing	.20* (4.56)	-	
3. Accuracy Prosody	.38*** (>100)	.26** (40.10)	-
4. Music Training	.12 (0.44)	.02 (0.10)	-.03 (0.10)

Note. Coefficient (BF_{10}). Spearman's rho correlations involving music training, Pearson's correlations between accuracy variables. *df* = 157. $p<.05$*, $p<.01$**, $p<.001$***.

2.3.3 Valence and Arousal Perceptions

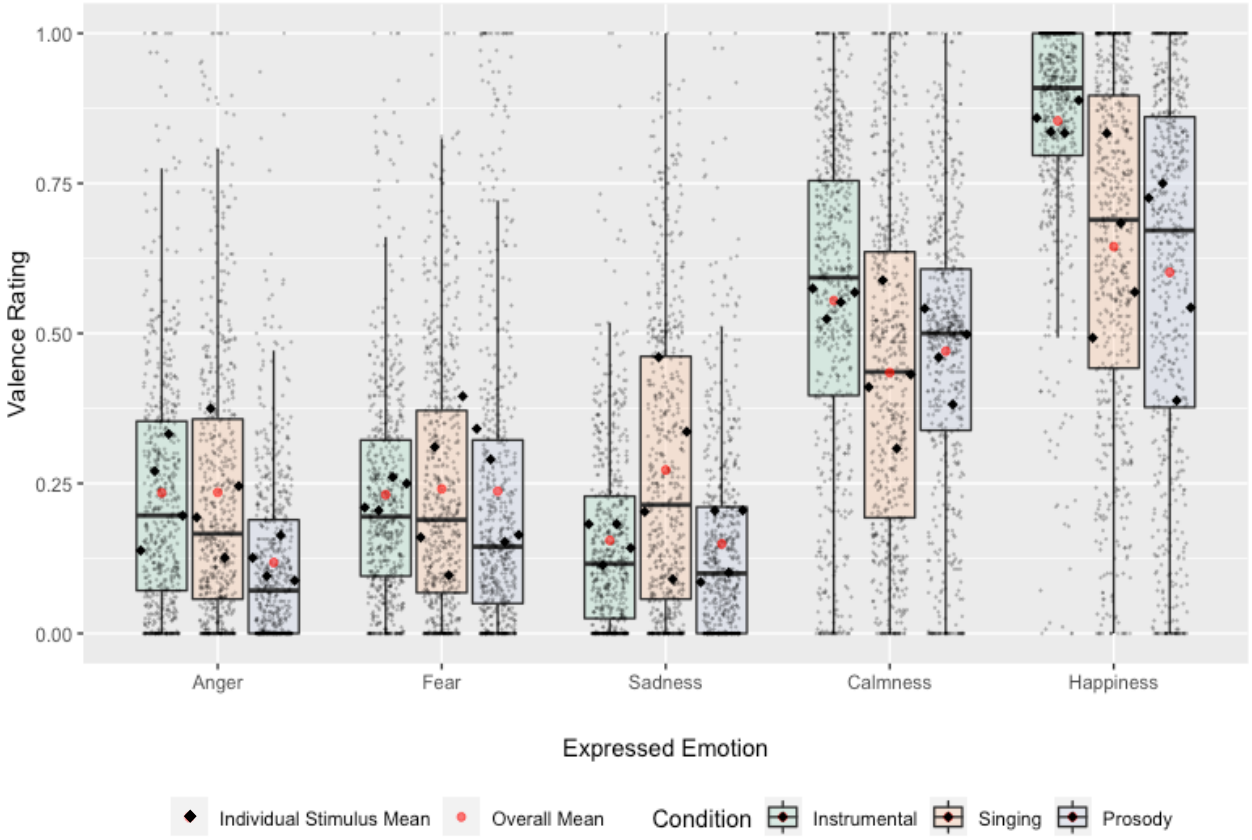
2.3.3.1 Valence.

Valence perceptions were relatively similar between conditions, although more between-condition variance was apparent for calmness and happiness, as reflected in Figure 2.4.[10]

[9] All correlations between accuracy scores and instrumental/theory training separately were also non-significant.

[10] Raw mean arousal and valence ratings can be seen in Appendix I.

Figure 2.4 – Distribution of Raw Valence Ratings by Emotion and Condition



Note. Smallest dots = per-trial observations. Boxplots display medians, 25th and 75th percentiles, and minimum/maximum values. Means = overall means and means for individual stimuli within each emotion/condition.

A LMM was conducted to analyse differences in valence ratings by condition and emotion, while accounting for level of music training. The LMM had random intercepts for participant and stimuli, and a by-participant random slope for emotion. The by-participant random slope for emotion significantly improved model fit (c^2 (14) = 1008.70, p<.001). The random intercepts for participant (c^2 (1) = 210.59, p<.001) and stimuli (c^2 (1) = 1457.90, p<.001) also improved model fit.

There was a significant main effect for condition, although the BF provided anecdotal evidence against this effect (F (2, 60.1) = 5.82, p =.005, BF_{10} = 0.31). On average,

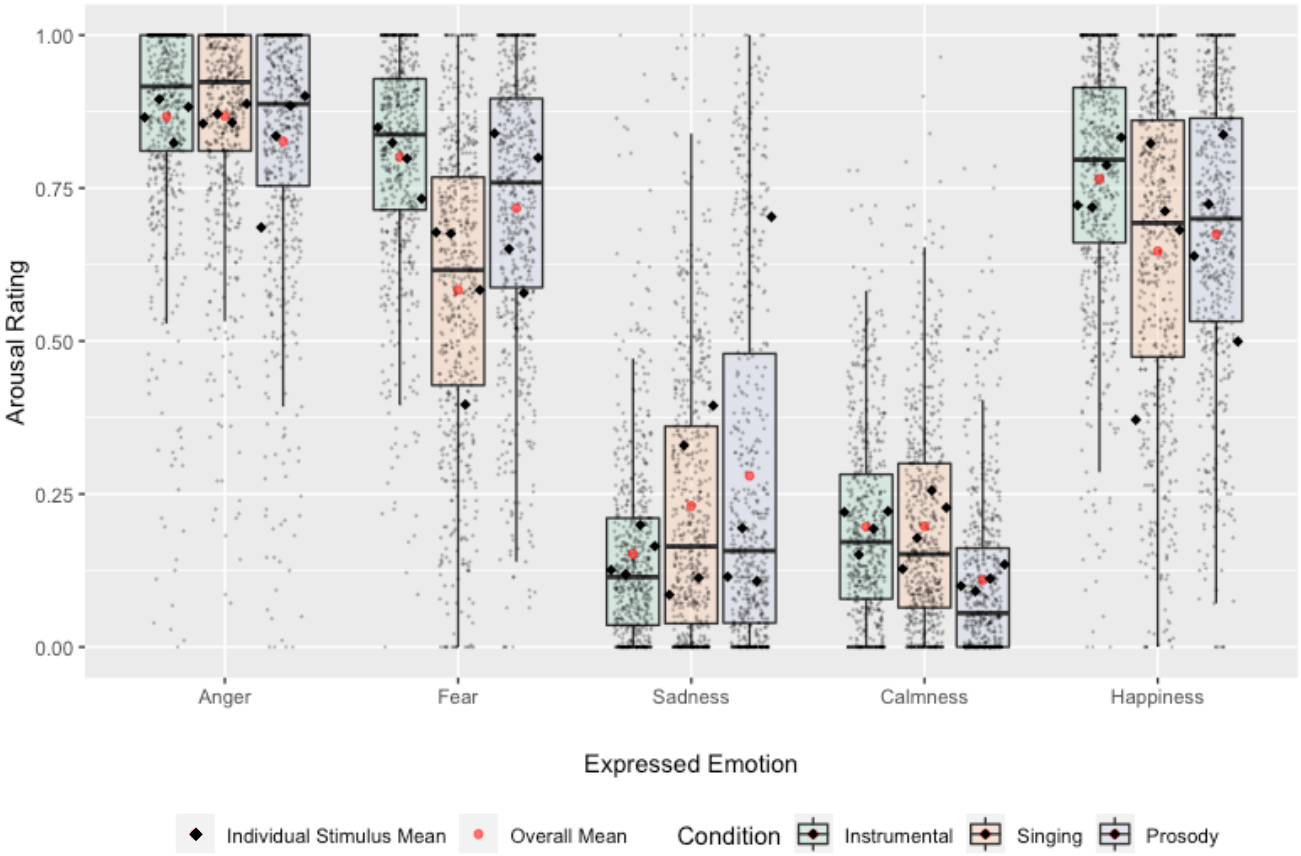
instrumental music stimuli (M = 0.41, SE = 0.02) were higher in valence than prosody stimuli (M = 0.31, SE = 0.02, p = .002, BF_{10} = 0.94), while there was strong evidence against a difference between other conditions (ps >.05, BF_{10s} < 0.30). There was a significant main effect for emotion (F (4, 70.2) = 67.08, p<.001, BF_{10} >100). All emotion pairs differed in valence (ps<.001, BF_{10s} >100) except for anger (M = 0.20, SE = 0.04) and fear (M = 0.24, SE = 0.02), anger and sadness (M = 0.19, SE = 0.03), and fear and sadness (ps>.05, BF_{10s} <0.10).

There was a significant interaction between emotion and condition, although the BF did not provide substantial evidence either for or against this effect (F (8, 60.1) = 2.77, p =.01, BF_{10} = 0.64). Perceived valence for happiness was higher on average for instrumental music (M = 0.85, SE = 0.04) compared to prosody (M = 0.60, SE = 0.04, p =.002, BF_{10} = 8.61) and singing stimuli (M = 0.65, SE = 0.04, p =.04, BF_{10} = 1.61).

2.3.3.2 Arousal.

There was congruence between conditions for arousal ratings. However, small differences were again apparent between conditions for certain emotions, while some individual stimuli (e.g., for sad prosody) did not align with average patterns (Figure 2.5).

Figure 2.5 – Distribution of Raw Arousal Ratings by Emotion and Condition



Note. Smallest dots = per-trial observations. Boxplots display medians, 25th and 75th percentiles, and minimum/maximum values. Means = overall means and means for individual stimuli within each emotion/condition.

A LMM analysis was conducted to analyse differences in arousal ratings by condition and emotion. The LMM had random intercepts for participant and stimuli, and by-participant random slopes for condition and emotion. The by-participant random slope for emotion significantly improved model fit (c^2 (15) = 1535.2, p<.001), as did the one for condition (c^2 (6) = 153.82, p<.001). The random intercepts for participant (c^2 (1) = 850.32, p<.001) and stimuli (c^2 (1) = 2731.70, p<.001) also improved model fit.

There was a significant main effect for emotion (F(4, 65.1) = 91.23, p<.001, BF_{10} >100). Most emotion pairs differed significantly in arousal, with anecdotal evidence for a

difference between anger (M = 0.85, SE = .032) and fear (M = 0.70, SE = .032, p = .003, BF_{10} = 2.79), moderate evidence for a difference between anger and happiness (M = 0.69, SE = .031, p = .002, BF_{10} = 3.76), and decisive evidence for a difference between anger and all other emotions (ps <.001, BF_{10s} >100). There was strong evidence for no difference in arousal between fear and happiness, and sadness (M = 0.22., SE = .031) and calmness (M = 0.17, SE = .031, ps>.05, BF_{10s} <.01). There was moderate evidence against an effect for condition, and anecdotal evidence against the interaction between condition and emotion (ps>.05, BF_{10s} <0.42).

2.3.4 Acoustic Feature Patterns

Acoustic feature levels for each stimulus were correlated with perceptions of valence and arousal, and the proportion of time each emotion was selected, to explore which features were related to participants' perceptions.

2.3.4.1 Valence and Arousal.

As Table 2.3 shows, there were strong positive correlations, across conditions, between mean loudness and arousal ratings. Tempo and arousal were also positively correlated across conditions. Brightness was positively correlated with arousal ratings for prosody and singing, but with valence ratings for instrumental music. The only prominent cross-condition pattern for valence was a positive correlation with mean pitch for instrumental music and singing stimuli. Pitch was also strongly related to perceptions for prosody, but this was for arousal and not valence.

All features were also entered into GLMMs to assess overall predictive strength of the feature set for arousal and valence perceptions. Arousal (M = 0.60, SD = 0.04) was better predicted by the acoustic features than valence (M = 0.27, SD = 0.08, p =.03; BF_{10} = 3.12), and this appeared consistent across conditions. Descriptively, valence was more strongly

predicted by the acoustic features for instrumental (R^2 = .34) and singing (R^2 = .28), relative to prosody stimuli (R^2 = .18).

Table 2.2 – Correlations Between Acoustic Features and Valence and Arousal Ratings

	Average Stimuli Ratings					
	Arousal			Valence		
Features	**Inst.**	**Sing**	**Pro.**	**Inst.**	**Sing**	**Pro.**
Loudness (mean)	.89*** (>100)	.92*** (>100)	.91*** (>100)	-.04 (0.48)	.04 (0.47)	-.23 (0.72)
Loudness (variation)	-.23 (0.69)	.49* (2.60)	.25 (0.90)	.02 (0.47)	-.01 (0.50)	-.60** (22.38)
Pitch (mean)	-.26 (0.76)	-.40^ (2.05)	.79*** (>100)	.49* (3.23)	.60** (26.42)	.02 (0.51)
Pitch (variation)	-.01 (0.47)	.17 (0.64)	-.60** (4.61)	.19 (0.61)	-.43^ (3.81)	-.14 (0.71)
Tempo/speech rate	.72*** (59.48)	.67*** (14.70)	.61** (6.04)	.44^ (2.37)	-.08 (0.49)	.13 (0.62)
Brightness	.22 (0.68)	.74*** (>100)	.53* (13.81)	.49* (2.73)	.17 (0.47)	-.44^ (2.76)
Marginal R^2	.59	.57	.64	.34	.28	.18

Note. Coefficient (BF_{10}). Data points = each stimulus (n = 20). Acoustic data = raw data. Spearman's correlations. p<.05*, p<.01**, p<.001***, p<.10^. Multi-condition correlations highlighted. Inst = instrumental; Sing = singing; Pro = prosody.

2.3.4.2 Emotion Perceptions.

Acoustic features were then assessed in relation to emotion perceptions, the results of which are in Table 2.4. As with arousal, loudness and tempo showed the most consistent relationships between feature levels and emotion perceptions. Other relationships were more sporadic - positive relationships between mean pitch and happiness perceptions across conditions, and brightness was associated with anger and calmness perceptions for prosody and singing stimuli. Pitch variables (both mean and variation) were more consistently related to emotion perceptions for prosody and singing stimuli, compared to instrumental stimuli. While the overall predictive strength of acoustic features for instrumental music (M = 0.45,

SD = 0.06) and prosody (M = 0.50, SD = 0.1), and instrumental music and singing (M = 0.37, SD = 0.16), did not differ (ps >.05, BF_{10s} <1), average predictive accuracy was higher for prosody than singing stimuli (t (4) = 3.30, p =.03, BF_{10} = 3.21).

Table 2.3 - Correlations Between Acoustic Features and the Proportion of Time Each Emotion was Selected, and Predictive Value of Whole Set of Acoustic Features, By Condition.

	Emotion Selection														
	Anger			**Fear**			**Sadness**			**Happiness**			**Calmness**		
Features	**Inst.**	**Sing**	**Pro.**	**Inst.**	**Sing**	**Pro.**	**Inst.**	**Sing**	**Pro.**	**Inst.**	**Sing**	**Pro.**	**Inst.**	**Sing**	**Pro.**
Loudness (mean)	.78 *** (>100)	.73 *** >100	.61 ** (10.46)	.65 ** (10.08)	.06 (0.47)	.23 (0.72)	-.60 ** (7.14)	-.67 ** (52.21)	-.54 * (5.70)	-.07 (0.54)	.08 (0.49)	.17 (0.59)	-.87 *** (>100)	-.68 *** (52.54)	-.80 *** (>100)
Loudness (variation)	-.16 (0.58)	.24 (0.72)	.50 * (3.81)	-.19 (0.60)	-.17 (0.63)	-.15 (0.65)	.21 (0.68)	-.13 (0.55)	-.15 (0.56)	-.10 (0.50)	-.01 (0.27)	-.70 *** (53.58)	.02 (0.48)	-.43 ^ (2.47)	-.37 ^ (1.30)
Pitch (mean)	-.19 (0.64)	-.59 ** (8.83)	.29 (0.71)	-.31 (0.94)	-.35 (1.21)	.68 *** (52.74)	-.02 (0.47)	.38 ^ (1.40)	-.24 (0.57)	.41 ^ (1.29)	.37 ^ (1.37)	.47 * (2.21)	.24 (0.79)	.70 ** (48.0)	-.73 *** (83.73)
Pitch (variation)	-.02 (0.48)	.46 * (2.38)	-.09 (0.56)	-.02 (0.52)	.37 (1.29)	-.67 ** (8.80)	-.25 (0.72)	-.29 (0.93)	-.01 (0.50)	.30 (1.02)	-.44 * (1.56)	-.44 ^ (2.58)	-.06 (0.47)	-.58 ** (9.27)	.49 * (2.26)
Tempo	.54 * (4.04)	.67 ** (18.07)	.24 (0.73)	.02 (0.47)	.60 ** (8.93)	.66 ** (12.49)	-.78 ** (>100)	-.58 ** (10.34)	-.29 (0.98)	.40 ^ (1.86)	-.17 (0.47)	.54 * (4.71)	-.59 ** (8.05)	-.80 *** (>100)	-.57 ** (5.22)
Brightness	.19 (0.64)	.71 *** (79.19)	.70 *** (37.88)	.01 (0.48)	.18 (0.63)	.01 (0.54)	-.24 (0.79)	-.72 *** (71.56)	-.31 (1.05)	.31 (1.00)	.05 (0.28)	-.30 (1.27)	-.23 (0.73)	-.71 *** (54.68)	-.62 ** (8.50)
Marginal R^2	.55	.62	.63	.45	.29	.46	.39	.30	.38	.44	.22	.46	.42	.44	.59

Note. Coefficient (BF_{10}). Data points = each stimulus (n = 20). Spearman's correlations. Data = z-scores. p<.05*, p<.01**, p<.001***, p<.10^. Multi-condition correlations highlighted. R^2 = marginal value for acoustic features (not including variance explained by random variance between participants).

2.4 Discussion

The present chapter found that adults' emotion recognition accuracy was similar for instrumental music and vocal prosody and correlated between all audio conditions. However, singing was less accurately recognised overall, and there were some condition-specific accuracy patterns for some emotions such as anger (more salient in prosody) and happiness (more salient in instrumental music). The present findings also indicated similarities in participants' perceptions of valence, and especially arousal, between audio conditions. Acoustic features were most strongly related to arousal perceptions, relative to valence, while arousal-related acoustic features were also most consistently related to emotion perceptions across conditions. However, differences in emotion recognition patterns and associated acoustic features for certain emotions may not support discrete emotion theories that propose direct mapping of expressive acoustic features to emotion categories across audio conditions (e.g., Juslin, 2018). Instead, they more closely align with dimensional theories that centralise arousal and valence within similarities in emotion expression and perception for musical and vocal stimuli (e.g., Cespedes-Guevara & Eerola, 2018). However, as is discussed, findings may indicate some condition-specific processes regarding how arousal and valence relate to emotion recognition.

In line with hypotheses, instrumental music and prosody showed similar overall recognition accuracy and recognition for each condition (including singing) were significantly correlated. This aligns with past research showing similar overall emotion recognition accuracy for musical and vocal stimuli in adults and children (Heaton & Allgood, 2015; Nordstrom & Laukka, 2019; Scherer et al., 2017; Vidas et al., 2018). There was also no correlation with years of music training for any condition, while acoustic features showed a similar level of predictive power for instrumental and prosody stimuli. This may suggest

alignment between conditions based on expressive features, rather than individual-level abilities such as music training. However, music-related experience was assessed using an ad hoc index based on years of formal training. While informative, this approach may have overlooked important dimensions such as passive musical engagement or self-guided experience. Standardised measures such as the Goldsmiths Musical Sophistication Index (GOLD-MSI - Müllensiefen et al., 2014) could be adopted in future research to more comprehensively capture individual differences in musical expertise and listening behaviours, strengthening assertions regarding the level to which musical and vocal emotions are related based on expressive features. Regardless, these findings may support theorised functional similarities between musical and vocal emotions (Juslin, 2013, 2018), linked to a shared evolutionary origin in early music-like social communication via the voice (Brown, 2017; Clark et al., 2015).

However, while some emotions (fear, calmness) also displayed similar emotion recognition accuracy levels between instrumental and prosody conditions, other emotions differed. This may question the idea that music and voice are fundamentally linked in their expressive properties and perceptual patterns at the level of discrete emotions (Juslin & Sloboda, 2011; Scherer, 1995). First, happiness (and to some extent sadness) was easier to recognise in instrumental music relative to other conditions. For happiness, this aligns with past research, with adults able to identify instrumental happiness with high accuracy (Lima et al., 2011; Veiellard et al., 2008), but less confident in identifying neutral or positive prosodic emotions relative to negative ones (Lausen & Hammerschmidt, 2020). For sadness, some past research has indicated similarly high accuracy levels for sad prosody and instrumental stimuli (Vidas et al., 2018). However, in those studies that included further low arousal emotions (such as calmness), sad prosody was more difficult to recognise, with levels close to that seen in the present study (e.g., Lausen & Hammerschmidt, 2020). Conversely, anger was more

accurately recognised in prosody relative to instrumental music - aligning with past research (Mohn et al., 2011; Vidas et al., 2018). These between-condition differences question the core tenet of discrete theories of musical and vocal emotion expression (Juslin & Sloboda, 2011; Scherer, 1995) – that perceptual links between music and voice are built around innate expressive similarities between emotion categories.

From a dimensional perspective, music and voice are consistently perceived as emotions, but these perceptions are substantiated by non-discrete variations in valence and arousal (e.g., Russell, 2003). Accurate emotion recognition also requires top-down understanding of emotion concepts and in the case of music, culture-specific cues such as major/minor mode (Cespedes-Guevara & Eerola, 2018). In the present experiment, despite differences in emotion recognition accuracy for certain emotions, arousal and valence judgements generally aligned across all audio conditions. This was even the case for singing stimuli, despite lower overall recognition accuracy for this audio condition. This similarity in arousal/valence perceptions was underpinned by similar overall predictive strength of acoustic features on arousal perceptions (and valence perceptions relative to instrumental music), and similarities in the specific acoustic features related to arousal perceptions (loudness and tempo). Thus, despite relative ambiguity in terms of expressed emotion, singing stimuli appears to align with other audio conditions in relation to affective dimensions. This aligns with past audio emotion recognition research indicating that ambiguous instrumental and prosody stimuli are more accurately perceived in terms of dimensions rather than emotion categories (Eerola & Vuoskoski, 2011; Holz et al., 2021). This relative similarity in perceptions of arousal/valence dimensions between conditions, despite some differences in emotion recognition, aligns with the primacy afforded to fundamental affective dimensions within the dimensional perspective (Cespedes-Guevara & Eerola, 2018; Laukka et al., 2005).

As hypothesised, between-condition alignment was particularly pronounced for arousal. There was confusion between low arousal emotions calmness and sadness across audio conditions, and no evidence for a between-condition difference in how emotions were perceived in terms of arousal. Although the confusion of negative emotions sadness and fear for prosody (and to a lesser extent singing) does not align with this overall pattern, one sad prosody stimulus was perceived as far higher in arousal than all other sad stimuli (see Figure 2.5). Similarly, one fear singing stimulus was perceived as lower in arousal than other fear stimuli. This could mean that fear and sadness were confused for these specific stimuli due to arousal similarity. For valence, perceptions differed to some degree between conditions, with instrumental happiness perceived as higher in valence relative to other audio conditions. This may have made it easier to distinguish from negative emotions, perhaps due to music-specific valence indicators such as musical mode (Gomez & Danuser, 2007). Indeed, happiness was more often erroneously perceived as fear for prosody stimuli relative to instrumental stimuli. Relative ambiguity in valence perception for prosody, in the absence of similar salient valence indicators (Sauter et al., 2010), could also underpin general biases towards perceiving negative emotions in past research (Nelson & Russell, 2011).

Perceptual patterns for singing stimuli may add further support for the cross-condition prominence of arousal within perceptual patterns. Past research has indicated that valence may operate as an affective dimension that refines more automatic arousal perceptions (Holz et al., 2021). In the present chapter, singing stimuli were closely matched to prosody in terms of length and complexity, but included salient musical mode cues that can effectively distinguish negative from positive emotions (Gomez & Danuser, 2007). Considering the similarities between audio conditions in these proposedly more 'automatic' arousal perceptions in the present chapter, one would expect that this salient valence indicator would align perceptual patterns for singing stimuli most closely with instrumental music. However,

while overall accuracy was lowest for singing stimuli, patterns of recognition/confusion for singing aligned most closely with those for prosody. For example, singing did not show the same pattern of confusion between anger and fear as instrumental music, but did display confusion between happiness and fear, and sadness and fear, in line with prosody. Further, the positive correlation between recognition accuracy for singing and prosody was stronger than the one between singing and instrumental music, suggesting closer alignment between the ability to recognise emotions in these vocal forms of emotion expression. This appears to support the idea that the physiological constraints relevant to vocal forms of emotion expression, and closely linked to expressed arousal, are more influential during emotion recognition for singing stimuli than musical conventions related to valence (Banse & Scherer, 1996; Goudbeek, 2010; Holz et al., 2021).

Some findings in the present chapter align less closely with this dimensional perspective, however, and may point to possible differences between musical and vocal stimuli in how arousal, valence, and emotions are processed. For example, there was a prominent pattern of confusion between anger and fear for instrumental music only, consistent with past research (Micallef Grimaud & Eerola, 2021; Sauter et al., 2010; Vidas et al., 2018). Given similarities between conditions in valence and arousal perceptions for these emotions in the present chapter, one would expect similar patterns of confusion across audio conditions. This discrepancy can be linked to possible differences between conditions in how arousal, valence, and emotion categories are processed during emotion recognition. Indeed, some acoustic features can communicate vocal anger independent of perceived arousal (Bänziger et al., 2015). This aligns with an integrated theoretical perspective, within which arousal and valence underpin and substantiate emotion perceptions, but certain expressive features can also relate more directly with emotion perceptions (see Chapter 1, section 1.6.3 for discussion). Indeed, neuroimaging research suggests overlapping but distinct perceptual

systems for arousal/valence and emotion categories during audio emotion perception (Giordano et al., 2022). Given the unique early developmental role of prosody in directing attention to socially meaningful information (Trainor et al., 2000) and the adaptive significance of quickly interpreting vocal anger signals (Scherer, 2003; Buss, 2005), the specificity of these direct links between expressive features and anger perception to vocal stimuli are plausible. This could explain the confusion between anger and fear for instrumental stimuli only (and associated higher recognition accuracy for angry prosody), in the present chapter.

In line with perceptual patterns, the associations between acoustic feature patterns and participants' perceptions of arousal/valence indicated similarities and differences between audio conditions. Across conditions, the set of acoustic features more strongly predicted participants' perceptions of arousal relative to valence, while prediction strength for arousal was similar between audio conditions. Although the present acoustic feature set was small, limiting the strength of inferences, this aligns with past research (Bänziger et al., 2015; Scherer et al., 2015; Scherer et al., 2017; Weninger et al., 2013) and may support proposed arousal salience in the expressive links between musical and vocal emotions (Cespedes-Guevara & Eerola, 2018). Further, arousal perceptions were related to loudness and tempo/speech rate across conditions, supporting the proposed importance of these features to arousal communication in music and voice (Bänziger et al., 2015; Llie & Thompson, 2006; Weninger et al., 2013). Pitch-related features and brightness were also correlated with perceived arousal for prosody and singing conditions. Conversely, pitch-related variables were related to perceived valence for instrumental music, aligning with past research indicating differing relationships between mean pitch and valence perceptions for music and voice (Llie & Thompson, 2006). Although only descriptively, instrumental and singing valence perceptions were more strongly predicted by the feature set compared to prosody, in

line with previous research (Weninger et al., 2013). Considered alongside perceptual patterns, this may support theoretical and empirical research highlighting arousal as a prominent condition-general affective dimension underpinning audio emotion perception (Cespedes-Guevara & Eerola, 2018; Weninger et al., 2013).

Associations between emotion perceptions and acoustic features also indicated some similarities and differences between audio conditions. In line with proposed arousal-prominence, the most consistent cross-condition correlations were for acoustic features loudness and tempo/speech rate, which were also related to arousal across conditions. However, there were many condition-specific patterns, including pitch variables' associations with the perception of some emotions for prosody and singing stimuli, but less so for instrumental stimuli. Accordingly, acoustic correlations don't appear to align with theories that claim that acoustic features align with discrete emotion categories consistently across audio conditions (Juslin & Laukka, 2003). However, the overall average predictive strength of acoustic features on participants' emotion perceptions did not differ between instrumental and prosody conditions, suggesting similar importance of the selected acoustic features across conditions. Recent research has indicated that musical aptitude – the ability to perceive differences in acoustic qualities in stimuli – predicts individual differences in musical and vocal emotion recognition accuracy (Jansen et al., 2023; Vigl et al., 2024). Thus, it may be that while some expressive acoustic similarities exist between audio conditions (particularly those related to arousal – Cespedes-Guevara & Eerola, 2018), the general ability to comprehend variations in acoustic features more strongly underpins overall emotion recognition similarities between musical and vocal stimuli seen in the present chapter and past research (Nordstrom & Laukka, 2019; Vidas et al., 2018). This supports shared mechanisms between audio conditions at the acoustic-perceptual level – aligning with evolutionary theories of musical and vocal emotions (Brown, 2017; Clark et al., 2015) –

while allowing for certain condition-specific expressive properties (e.g., musical mode) and processing mechanisms (e.g., for anger recognition in prosody).

2.4.1 Limitations and Future Directions

The present chapter has a range of limitations. First, acoustic analyses were purely correlational and limited by the size of the stimulus set and feature set. Therefore, correlations with individual features did not account for possible interactions with other features. Although many of the associations between acoustic features and emotion perceptions aligned with past research, inferences should not be over-generalised. Further, given the limited size of the stimuli set, acoustic analyses related to the proportion of time each emotion was selected, rather than emotion recognition accuracy. Future research with more stimuli could explore how acoustic features relate to recognition accuracy, and how these patterns compare across audio conditions, in line with past research for vocal (e.g., Lima et al., 2016) and musical stimuli (e.g., Laukka et al., 2013b). Further, inferences regarding the strength of prediction by the acoustic feature set are specific to the set of features selected. It is likely that further features, particularly condition-specific expressive devices such as mode for music stimuli, and pitch direction for vocal stimuli (Juslin & Laukka, 2003), would have increased the strength of prediction for perceived valence.

Second, some of the inferences drawn would be strengthened by consideration of further emotion categories. For example, past research has suggested that a wide range of positive emotions can be communicated via vocal prosody, including awe, pride, and relief (Cowen et al., 2019). Accordingly, it may be that limiting positive emotions to happiness and calmness represented an oversimplification of vocal emotion expression, which may have influenced the relative difficulty recognising prosodic happiness in the present chapter. Although some past music research has attempted to explore emotion recognition for a wider set of emotions, such as love and pride (Micallef Grimaud & Eerola, 2021; Vidas et al.,

2018), studies have generally explored a relatively narrow set of musical emotions. Future cross-condition research could explore a wider range of emotion categories. Considering arousal/valence judgements alongside emotion recognition may facilitate exploration of a wider range of emotions, as even emotions with low recognition accuracy can be interpreted in terms of perceived arousal and valence levels, to inform understanding of cross-condition similarities and differences.

Finally, comparisons between discrete and dimensional theories of musical and vocal emotion expression/perception are limited by the stimulus selection approach. Stimuli were selected based on intended expression of an emotion category. Therefore, inferences regarding similarities between conditions in terms of arousal and valence perceptions were based on their levels within these emotion categories. Although this aligns with evidence regarding close alignment between dimensional and discrete perceptions of emotions for music and voice (Eerola & Vuoskoski, 2011; Sauter et al., 2010), this limits inferences regarding perceptions of these dimensions independent of their alignment with specific emotions. It may be that selecting stimuli based on their expressed arousal/valence properties and collecting emotion perception data could provide unique insights regarding similarities and differences between audio conditions, particularly when considering how these affective dimensions directly relate to/predict emotion recognition patterns (see Vuoskoski & Eerola, 2011).

2.4.2 Conclusion

Overall, perceptual patterns support elements of a dimensional theoretical perspective that considers arousal and valence as important cross-condition affective dimensions that underpin and substantiate emotion perceptions for musical and vocal stimuli (Cespedes-Guevara & Eerola, 2018). It appears that arousal may be particularly important for audio emotion recognition, as indicated by close alignment in perceptual patterns between singing

and prosody (possibly due to shared physiological constraints), similar arousal perceptions across audio conditions, and the predictive value of acoustic features for arousal relative to valence perceptions. This aligns with past vocal and musical research (Holz et al., 2021; Weninger et al., 2013), but differs from facial expression research, where valence appears the most salient affective feature for both children and adults' emotion perceptions (Woodward et al., 2022). Further, there may be some more direct links between expressive features and the perception of certain vocal emotions (Bänziger et al., 2015; Giordano et al., 2021), and differences between audio conditions regarding the salience of valence (Weninger et al., 2013). These differing expressive and perceptual processes could have given rise to the present condition-specific recognition patterns, such as the salience of anger for prosody, and happiness for instrumental music. Accordingly, audio emotion recognition research may be most informative if considering elements of both dimensional and discrete emotion theories. This integrated approach may be especially fruitful for developmental research, where developing understanding of arousal and valence properties may be key to improvements in emotion recognition accuracy (Kragness et al., 2021; Nelson & Russell, 2011; Widen, 2013).

3. Emotion Recognition in Instrumental Music, Singing, and Vocal Prosody in Typically Developing Children

3.1 Relationship to Previous Chapters

Chapter 2 examined adults' emotion recognition accuracy and perceptions of underpinning valence and arousal dimensions for instrumental, vocal prosody, and singing stimuli. A set of acoustic features were also extracted and analysed in relation to perceptual patterns. Findings revealed positive associations between recognition accuracy for audio conditions, and similarities in adults' valence, and especially arousal, perceptions. However, there were also some condition-specific accuracy patterns for certain emotions, with anger easier to recognise in prosody stimuli, happiness (and sadness to some extent) more salient in instrumental music, and singing stimuli most difficult to recognise overall. Across conditions, acoustic features were most strongly and consistently related to perceived arousal, relative to valence, and these arousal-related acoustic features were the most consistent cross-condition features related to emotion perceptions. Findings supported theoretical cross-condition primacy of the fundamental affective dimension arousal during audio emotion recognition (Cespedes-Guevara & Eerola, 2018; Holz et al., 2021) but suggested possible differences between conditions in the salience of valence and associations with acoustic features.

Cross-condition research exploring the typical development of emotion recognition in these audio stimuli remains scarce. Given rapid development of emotion recognition ability during childhood (Grosbras et al., 2018), developmental research is uniquely positioned to highlight the extent to which emotion recognition patterns between musical and vocal stimuli align. Further, considering the proposed importance of affective dimensions arousal and valence to developing emotion recognition accuracy (Nelson & Russell, 2011; Widen, 2013), such research is well placed to provide theoretical insights regarding the nature of links

between audio conditions. Accordingly, the present chapter examined emotion recognition for vocal and musical stimuli across a key period of emotional development. Arousal and valence (normative adult ratings from Chapter 2) were analysed as stimulus-level features in relation to emotion perceptions.

3.2 Introduction

Vocal prosody emotion recognition is a key developmental skill – facilitating social interaction and positive socio-emotional adjustment (Denham, 1998; Saarni, 1999). Past research has demonstrated condition-general emotion processing mechanisms at the acoustic-perceptual level (Vigl et al., 2024) and reliance on variations in similar acoustic features to express emotions (Grandjean et al., 2006; Juslin & Laukka, 2003; Eerola et al., 2013; Scherer et al., 2015). Further, neural mechanisms involved in social cognition are important for emotion processing of both musical and vocal stimuli (Escoffier et al., 2013; Lima et al., 2016; Vigl et al., 2024). These cross-condition similarities are reflected in correlations between adults' emotion recognition accuracy for vocal and musical stimuli (both instrumental and singing - Chapter 2; Laukka & Juslin, 2007; Nordstrom & Laukka, 2019). Despite these similarities, there is disagreement regarding the extent to which similar emotion recognition patterns for music and voice relate to shared expressive acoustic features tied to discrete emotion categories (e.g., Juslin & Laukka, 2003), or to more basic affective dimensions arousal and valence (e.g., Cespedes-Guevara & Eerola, 2018). In the latter dimensional perspective, these more basic affective properties are integrated and categorised based on conceptual knowledge of emotion categories, facilitating accurate emotion recognition (Barrett, 2017; Russell, 2003). Indeed, musical and vocal stimuli appear particularly closely related in relation to perceived arousal (Bänziger et al., 2015; Chapter 2; Llie & Thompson, 2006; Weninger et al., 2013), although evidence suggests that emotions can also be communicated independent of these dimensions (Bänziger et al., 2015; Giordano

et al., 2021). One avenue through which to better understand similarities and differences between musical and vocal emotion recognition is through their typically developing trajectories. When considered developmentally, the importance of affective dimensions arousal and valence for emotion recognition becomes apparent.

3.2.1 Emotion Recognition Development – From Broad to Differentiated

Emotion recognition development involves multiple stages. Evidence indicates that before they develop expressive language abilities, children can distinguish some emotions based on their perceptual properties and link them to specific situations/outcomes (Ruba et al., 2018; Ruba & Repacholi, 2020). This suggests development of a functional understanding of specific emotion categories before proposed development of the ability to explicitly label emotion expressions. Following this stage, some researchers have conceptualised emotion recognition development in terms of developing understanding of fundamental dimensions arousal and valence (Widen, 2013). An increasingly differentiated comprehension of emotions also involves developing conceptual understanding of emotion categories – allowing children to organise their knowledge relating to specific emotion categories (Barrett, 2006). Indeed, valence and arousal dimensions strongly predict adults' emotion perceptions for both vocal and musical stimuli, suggesting that these dimensions underpin emotion recognition patterns (Eerola & Vuoskoski, 2011; Sauter et al., 2010). For facial stimuli, evidence indicates that valence is particularly important to early perceptions of emotion stimuli, while the role of conceptual understanding of emotion categories increases with age (Woodward et al., 2022).

Although the shift from broad dimensional understanding to more advanced emotion recognition ability may be relevant for audio stimuli (Nelson & Russell, 2011), little research has examined this directly. It may be the case that arousal has an increasingly prominent role for audio relative to visual stimuli. Indeed, arousal is accurately perceived in vocal stimuli

across different levels of expressed intensity, while valence perception is accurate only at certain intensity levels (Holz et al., 2021). This aligns with the proposed relative importance of arousal within musical emotion expression, possibly drawing on similar processing mechanisms to the voice (Cespedes-Guevara & Eerola, 2018). Indeed, acoustic features such as loudness and tempo are positively associated with perceived arousal in both music and voice, while arousal is more strongly predicted by acoustic features compared to valence in both stimulus types (Bänziger et al., 2015; Chapter 2; Llie & Thompson, 2006; Weninger et al., 2013). Accordingly, it may be that arousal-based understanding (i.e., comprehension of high versus low expressed energy levels) is particularly important to audio emotion recognition abilities, and to how these abilities develop.

3.2.2 Typical Development of Emotion Recognition in Vocal Prosody, Instrumental Music, and Singing Stimuli

3.2.2.1 Vocal Prosody.

Most research on the typical development of emotion recognition involves facial expressions. While findings vary based on methodological differences, the research base generally suggests substantial development before 10 years, and some continued development through adolescence for more challenging recognition tasks (e.g., with stimuli of differing intensities – Chronaki et al., 2015b; Gao & Maurer, 2010; Nelson & Russell, 2011).

Compared to facial expressions, there is limited research on prosody emotion recognition development. While the evidence that does exist is mixed, emotion recognition patterns appear partially explainable via a) differences in the stimuli adopted and b) increasing sensitivity to arousal and valence with age. In line with facial expression research, most prosody research indicates that there is rapid emotion recognition development between

pre-schoolers (4-5 years) and children (6-9 years), with some continued development into adolescence (Chronaki et al., 2015b; Grosbras et al., 2018). However, stimulus length appears to affect findings regarding accuracy for specific emotions, which may be linked to the ability of longer stimuli to express more acoustic information. For example, using single syllables as stimuli, Grosbras et al. (2018) found that happiness and sadness were easier to recognise than anger and fear, particularly early in development. This difficulty recognising anger and fear was underpinned by high levels of confusion between these emotions, perhaps due to their similarity in both arousal and valence (Russell, 2003). Conversely, when longer stimuli were adopted, anger and sadness were the earliest developing and most salient prosodic emotions (Nelson & Russell, 2011; Sauter et al., 2013; van Zonneveld et al., 2019; Zupan, 2015). This was the case even when a larger number of emotions and different methodological approaches (e.g., forced choice versus free labelling) were adopted (Nelson & Russell, 2011; Sauter et al., 2013). Given some of the proposed acoustic differences between anger and fear stimuli (e.g., pitch variation – Chapter 2; Juslin & Laukka, 2003), it may be that these longer stimuli allowed for interpretation of expressive differences between anger and fear, while this may have also contributed to the greater relative accuracy for sad stimuli (e.g., via interpretation of differences in speech rate).

Some of these patterns may also be attributed to increased sensitivity to the arousal and valence levels of vocal expressions with age. For example, Nelson and Russell (2011) suggested that biases towards negative emotions in their study with 3-5-year-old children indicated continued development of valence-based understanding of vocal emotions beyond 5-years. They also found that sadness was the best recognised vocal emotion, compared to anger, fear, and happiness (Nelson & Russell, 2011). This aligns with other research showing high accuracy for sadness early in development (Grosbras et al., 2018; Sauter et al., 2013). The early salience of sadness could also be linked to sensitivity to arousal, given that other

commonly presented emotions (happiness, anger, fear) are higher in perceived arousal (Chapter 2). This would align with research highlighting a stronger link between expressive acoustic features and perceived arousal relative to valence in the voice (Bäzinger et al., 2015), and greater relative salience of arousal properties in vocal stimuli (Holz et al., 2021). Indeed, arousal-related features loudness and speech rate are negatively associated with perceived vocal sadness in adults (Banziger et al., 2015; Chapter 2). Thus, it appears plausible that arousal properties are particularly important for prosodic emotion recognition early in development, while valence-based understanding has a less prominent role that continues to develop beyond 5-years – misaligned with the valence-based development pertinent to facial expressions (Nelson & Russell, 2011; Widen, 2013).

3.2.2.2 Instrumental Music.

Developmental research examining emotion recognition in instrumental stimuli is more limited than that for prosody. While research findings point to a similarly important early role for arousal-based understanding, differences relating to valence-based development also appear possible. While the ability to distinguish happiness and sadness at above chance level develops by 3-years (Franco et al., 2017; Stacho et al., 2013), the same ability for anger and fear may not develop until 6-8-years (Kratus, 1993; Nawrot, 2003). Adult levels appear to be reached for all these emotions by 11-years (Hunter et al., 2011), although response distributions match adult patterns much earlier (Nawrot, 2003). These patterns have been linked to developing sensitivity to arousal and valence properties and associated acoustic features. Specifically, early in development, children's emotion judgements are sensitive to musical tempo and loudness (Kragness et al., 2021; Mote, 2011). In line with prosody research, this may suggest early sensitivity to arousal, with strong links between tempo/loudness and perceived arousal in musical stimuli (Cespedes-Guevara & Eerola, 2018; Chapter 2; Llie & Thompson, 2006). Children's sensitivity to emotions in music based on

major/minor mode appears to develop later, around 6-8 years (Dalla-Bella et a., 2001). As musical mode is closely linked to perceived valence (Gomez & Danuser, 2007), this later developing sensitivity may align with an increase in the emotions that can be accurately recognised. Indeed, in a music production task, Kragness et al. (2021) found that 5-year-old children could manipulate loudness and tempo to distinguish emotions based on arousal, while the ability to do so based on valence was only apparent in 7-year-olds. Although valence-based development may also be important to vocal emotion recognition development (Nelson & Russell, 2011), the lack of a similar valence-based indicator to musical mode in the voice may lead to differences in how this aspect of perceptual emotion understanding develops.

3.2.2.3 Cross-Condition Research.

A cross-condition approach facilitates direct consideration of the commonalities between musical and vocal emotion recognition development. Two developmental studies have taken this approach using instrumental and prosody stimuli. Vidas et al. (2018) asked 60 6-11-year-old children to complete an emotion recognition task involving prosody, 30-second orchestral music clips, and affect bursts. Stimuli expressed happiness, sadness, fear, anger, and pride. They found that emotion recognition for each stimulus-type developed in parallel. The most pronounced similarities were between instrumental music and prosody, with recognition accuracy for prosody able to predict accuracy for music, independent of age and musical training. They also found that anger was more difficult than other emotions for children to recognise for music only, in line with past research with adults (Chapter 2). Using similar musical stimuli and affect bursts, Heaton and Allgood (2015) asked 5-10-year-old children to give emotion recognition judgements for happy, sad, and fearful stimuli. They also found a significant correlation between musical and vocal emotion recognition across all age groups, highlighting cross-condition similarities even early in development. Further,

musical stimuli were easier than vocal stimuli to recognise overall, supporting the idea that music is a highly expressive form of emotion stimuli (Juslin, 1997; 2018). The authors concluded that findings supported a possible 'cross-condition model of auditory emotion recognition' (Heaton & Allgood, 2015, p. 402).

3.2.2.4 Singing – A Developmental Bridge Between Instrumental Music and Vocal Prosody.

Past cross-condition studies have rarely included singing stimuli. However, incorporation of this form of audio stimuli may be informative for audio emotion recognition research, considering its unique links to both prosodic and instrumental forms of expression. Singing is uniquely related to vocal prosody via shared physiological constraints. This translates to close similarities between singing and prosody in the acoustic features that communicate emotion, particularly those related to arousal (Scherer et al., 2015; Scherer et al., 2017). Singing is also uniquely related to instrumental stimuli via shared musical conventions such as musical mode - strongly related to perceived valence (Gomez & Daunser, 2007). Singing may also offer unique insight on account of its role during development. Indeed, singing is more effective than speech in directing attention to the mouth, supporting various aspects of socio-emotional development (Alviar et al., 2023). This is unsurprising, given the 'music-like' nature of early mother-child interactions, underpinned by developing sensitivity to acoustic cues such as loudness and tempo (Flom & Bahrick, 2007; Trehub, 2001) and hypothesised as the basis of later developing vocal emotion recognition skills (Boone & Cunningham, 1998; Schubert & McPherson, 2015). Accordingly, considering singing alongside instrumental and prosody stimuli may be revealing regarding the nature of perceptual links between music and voice, and the ways in which recognition accuracy for these forms of audio stimuli develop.

While emotion recognition research involving singing stimuli is limited, findings appear closely aligned with instrumental music. Children as young as 3-years can distinguish between happiness and sadness in instrumental and singing stimuli with similar accuracy (Franco et al., 2017). Recognition accuracy for these emotions matures early in development, with 5-year-old children able to distinguish sung happiness and sadness comparably to adults (Morton & Trehub, 2007). Other research also found similarities between singing and instrumental conditions for emotion recognition of happiness, sadness, anger, and fear. However, instrumental sadness may be easier to recognise than singing for older participants (Dolgin & Anderson, 1990), while adults are also able to recognise emotions in vocal prosody more accurately than singing (Chapter 2; Livingstone & Russo, 2018). This may suggest closer alignment between emotion recognition accuracy for singing and other audio conditions early in development. It is unclear whether these perceptual patterns relate in the same way as other audio conditions to fundamental affective dimensions. However, children can expressively communicate happiness and sadness through singing from 4-years, relying primarily on expressive cues such as tempo and loudness (Adachi & Trehub, 1998). Further, loudness and tempo were associated with adults' arousal perceptions, and to perceptions of some emotions, in Chapter 2, suggesting a role for stimulus-level affective dimensions within developing emotion recognition for singing.

3.2.3 The Current Study

Considering its developmental importance (McPherson & Schubert, 2015), a cross-condition model of audio emotion recognition development (Heaton & Allgood, 2015) could be further explicated by incorporating singing stimuli into a developmental study. Further, the importance of comprehension of stimulus arousal and valence properties for both musical (Cespedes-Guevara & Eerola, 2018; Stacho et al., 2013) and vocal (Giordano et al., 2021; Holz et al., 2021; Nelson & Russell, 2011; Sauter et al., 2010) emotion recognition suggests

that consideration of these affective dimensions would be informative for any developmental conceptualisation. Such a conceptualisation should pay particular attention to early childhood (4-9 years), where audio emotion recognition development seems most rapid. Greater focus on these aspects could illuminate the tensions between discrete and dimensional theories discussed, while providing evidence regarding a possible cross-condition model of audio emotion recognition development (Heaton & Allgood, 2015). A stronger cross-condition model could open avenues for a more practical focus on music-based interventions, in line with positive associations between children's vocal emotion recognition, musical abilities, and socio-emotional dimensions (Blasco-Magraner et a., 2021; Neves et al., 2021). Accordingly, the current chapter explored emotion recognition in instrumental music, vocal prosody, and singing stimuli in TD 4-9-year-old children. Average adult-rated valence/arousal levels for each stimulus were analysed in relation to perceptual patterns, and any differences in these associations based on age and audio condition were explored.

3.2.3.1 Hypotheses.

1. Emotion recognition accuracy will be similar between audio conditions in younger children, but instrumental music and prosody stimuli will develop more quickly and be easier to recognise in older children, relative to singing stimuli. Children as young as three show similar accuracy for recognising happiness and sadness in singing and instrumental music (Franco et al., 2017), while young children can also recognise sadness in vocal prosody (Nelson & Russell, 2011; Grosbras et al., 2018). However, instrumental and prosodic emotion recognition develop rapidly across childhood (Chronaki et al., 2015b; Vidas et al., 2018; Kratus, 1993; Nawrot, 2003), and adults recognise prosodic emotions more accurately than sung emotions (Livingstone & Russo, 2018), suggesting less extensive development for singing relative to other audio stimuli.

2. In line with patterns with adults (Chapter 2), emotion recognition accuracy for instrumental, prosody, and singing stimuli will be positively associated.
3. Stimulus arousal will predict emotion perceptions in a consistent manner across conditions and age groups. Arousal is a salient and consistently perceived feature across vocal and musical stimuli (Cespedes-Guevara & Eerola, 2018; Holz et al., 2021). Acoustic cues such as loudness and tempo/speech rate are strongly associated with arousal in all three stimulus types (Ilie & Thompson, 2006; Bänziger et al., 2015; Weninger et al., 2013), and children can manipulate these cues to express emotion from as early as 4–5 years (Adachi & Trehub, 1998; Kragness et al., 2021). These consistent associations suggest that arousal will reliably predict emotion perceptions across conditions and developmental stages.
4. Relative to stimulus arousal, valence will display more variation between conditions and will show more age-related increase in its associations with emotion perceptions. While arousal-related cues such as loudness and tempo support early emotion perception across conditions, valence appears more variable and develops later. Sensitivity to musical mode, a key valence cue in music, emerges around age 6–8 (Dalla Bella et al., 2001), and valence understanding in prosody continues to develop beyond age 5 (Nelson & Russell, 2011; Widen, 2013). These differences suggest greater developmental change and between-condition variation in valence associations compared to arousal.

Accuracy patterns for specific emotions were analysed exploratively.

3.3 Methods

3.3.1 Participants

Ninety-five TD children aged 4-9 years (M = 7.21), and their parent/guardian were recruited via mainstream primary schools. There were 46 males and 49 females. Participants were not deaf or hard of hearing, spoke English fluently, and had no diagnosed developmental condition nor scored 'very high' for either internalising or externalising difficulties on the Strengths and Difficulties Questionnaire (SDQ – Goodman, 1997). These very high scores are most strongly related to a later diagnosis of a neurodevelopmental condition (Goodman et al., 2010). Participants with a standardised score on the British Vocabulary Picture Scale (BPVS – Dunn & Dunn, 2009) below 70 were also excluded, as this indicates a significant language difficulty. Sixteen participants had at least half a year of formal music training.

An a priori power analysis was conducted to estimate required sample size for the GLMM measuring differences in emotion recognition between age groups by condition and emotion. This was done using the *simr* R package (Green & MacLeod, 2016). Simulated data and effect sizes for fixed and random effects were estimated based on adult findings from Chapter 2, and past research on musical and vocal emotion recognition development (Vidas et al., 2018; Heaton & Allgood, 2015). Models were then simulated with 100 repetitions to assess power to detect a meaningful effect at a power of 0.8, with an alpha value of .05. The power analysis indicated that a minimum sample size of 26 children per age group (78 total) would be sufficient to detect a meaningful effect for each main or interaction effect. Accordingly, the sample of 95 children with 27 in the smallest age group, was sufficient. The study was granted ethical approval by the Cardiff University School of Psychology Research Ethics Committee.

3.3.2 Material and Procedure

Participants completed the tasks in-person, one-to-one with a researcher, in a quiet room free from distraction. Participants listened to the audio stimuli through JBL JR460 headphones with active noise cancellation. The emotion recognition task was created and presented on Gorilla experiment software (Anwyl-Irvine et al., 2020). Task instructions and response options were presented on-screen and verbally by the researcher, considering distinct developmental trajectories in the comprehension of written and verbal language (Wolf et al., 2019). Participants provided their responses verbally. The researcher controlled the pace of progression through the task, which lasted approximately 15 minutes. Parents/guardians also completed a set of questionnaires which took approximately 10-15 minutes to complete.

3.3.2.1 Emotion Recognition Task.

3.3.2.1.1 Materials.

For each audio condition (instrumental, singing, prosody), 20 stimuli were adopted - four for emotions anger, fear, sadness, happiness, and calmness. Stimuli were matched between conditions as far as possible. All stimuli were high intensity and 3-7 seconds in length, in line with the 'mental presence time' for audio stimuli (Argstatter, 2016). A random set of vocal prosody and singing stimuli were taken from the validated RAVDESS stimulus set (Livingstone & Russo, 2018). Instrumental piano pieces for each emotion were taken from a range of validated stimuli sets (Micallef Grimaud & Eerola, 2021; Sutcliffe et al., 2017; Vieillard et al., 2008). For more details on stimuli selection and properties, see Chapter 2 (section 2.2.2.1.1).

Stimuli were normalised as far as possible within and between conditions (see Appendix A for details). Hardware volume was also normalised across participants as far as

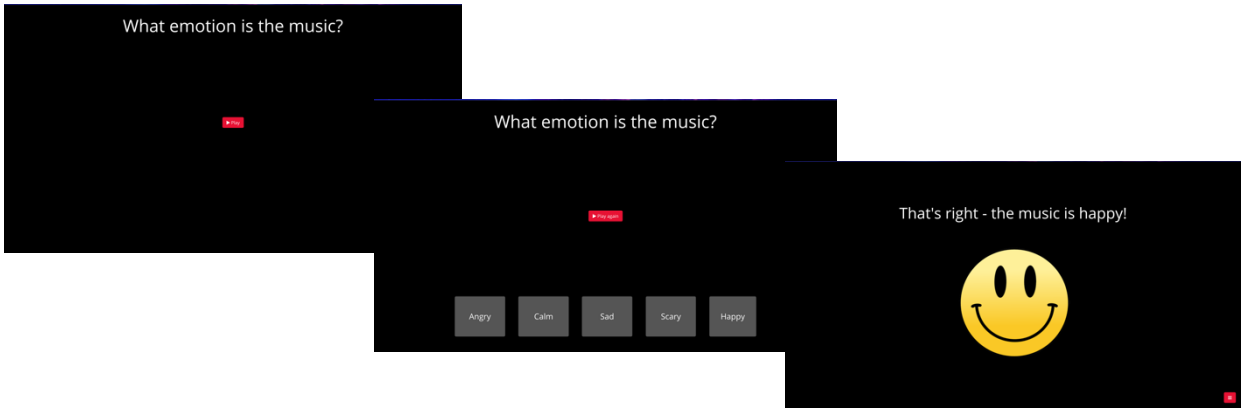
possible. However, an initial 'sound check' involving listening to a series of piano notes, peaking at 0.2 dBs above the maximum volume within the task, was conducted, to ensure the volume level was not uncomfortable. Participants that edited their sound level were recorded, and initial analyses were run both with and without these participants to ensure no difference in results.

3.3.2.1.2 Procedure.

Participants were first introduced to the overall aim, which involved helping Shaun the Sheep buy the farmer some new glasses by collecting 15 coins within a task about emotions. Participants were then told that they would hear a music clip, speaking voice, or singing voice, and to decide which emotion they thought the stimulus was expressing. Task trials involved listening to one of these clips and selecting from the given emotions within a forced-choice format.

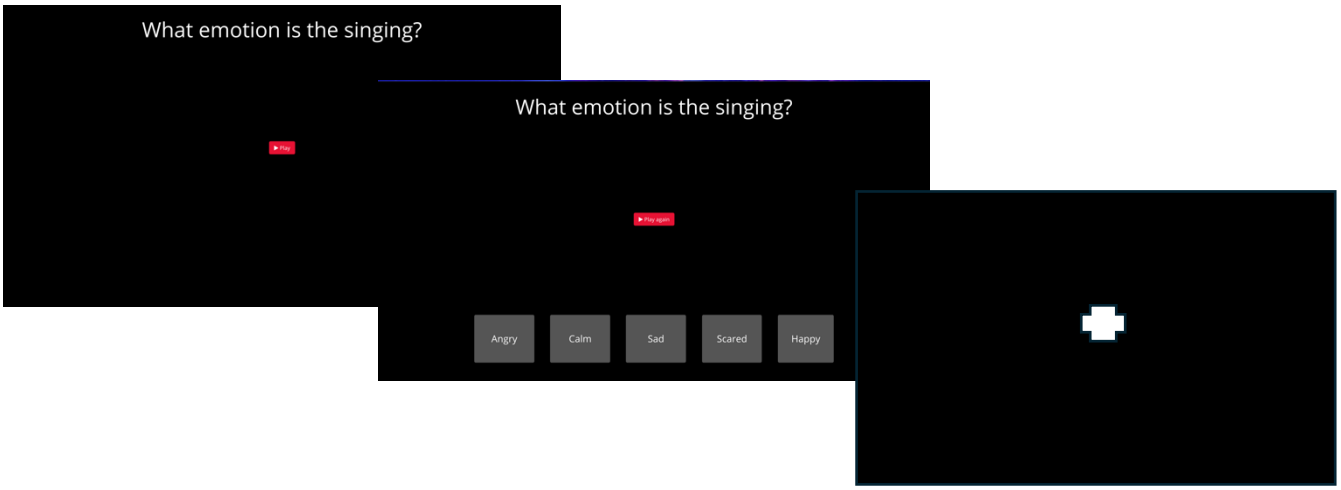
For each trial, the question 'what emotion is the music/voice/singing?' was presented on screen and read aloud by the experimenter before the music clip was triggered by clicking on the 'play' button. Once the clip finished, the response options appeared on screen and were read aloud by the experimenter. Children then provided their response verbally and the experimenter clicked on the relevant emotion word. There was an initial practice phase for each condition. Within this phase, participants heard one example stimulus for each emotion (differing from the stimuli within experimental trials) and received feedback regarding the intended emotion expression (Figure 3.1).

Figure 3.1 - Practice Trial for Musical Stimulus Expressing Happiness



After the practice phase, 20 experimental trials were presented for each condition (Figure 3.2). There was a three-second interval between trials during which the screen displayed a white cross on a black screen. After every fourth trial, participants earned a coin and were shown this on screen.

Figure 3.2 - Experimental Trial for Singing Stimulus



Once participants completed the whole task, they were taken through a series of screens showing all the coins they had earned, and Shaun the Sheep using the coins to replace the farmer's glasses. Stimulus conditions were presented separately and in a random order, and counterbalanced. Stimulus emotion within each condition were presented in an order A/order B quasi-random order (see Appendix L), counterbalanced across participants. Individual stimuli for each emotion were randomly allocated within this order, for each participant. For each participant, emotion recognition accuracy scores per emotion and condition were produced.

3.3.2.2 Stimulus Arousal and Valence.

Valence and arousal values were assigned to each audio stimulus, based on the mean of adult ratings from Chapter 2. In Chapter 2, adult participants rated the perceived valence (negative to positive feeling) and arousal (low to high energy) for each stimulus on a scale from 0 – 100, within an online experiment. Figure 2.4 and Figure 2.5 in Chapter 2 outline mean arousal and valence levels for each stimulus. These were z-scored within each condition to standardise them for analyses.

3.3.2.3 Parent Questionnaires.

Parents/guardians provided demographic information, indicated whether their child had a diagnosed developmental disorder and/or were deaf or hard of hearing. They also indicated their child's level of music training in terms of years of non-classroom music lessons, either instrumental or singing. The validated SDQ (Goodman, 1997) was completed to ensure included children did not have high levels of internalising or externalising difficulties which could suggest a developmental condition and confound results.

3.3.3 Statistical Analysis

In line with recent suggestions for psychiatric and psychological research (Keysers et al., 2020; Quintana & Williams, 2018), both frequentist and Bayesian analyses – in the form of Bayes Factors (BFs) – were included. BFs provide direct evidence for both the alternative and null hypotheses, important for the present research which aims to quantify both the presence and absence of effects. BFs were interpreted in line with Jeffery's (1998) specifications, outlined in Appendix D. Inferences were based on coherence between *p*-values and BFs (e.g., evidence against an effect included a *p*-value >.05 and a BF <1/3), and any incoherence was discussed. All analyses were conducted in R (R Core Team, 2022). Full details of statistical procedures, including assumption tests, specific R packages, and reporting, can be seen in Appendix E.

3.3.3.1 Frequentist Analyses.

3.3.3.1.1 Emotion Recognition Development

A GLMM with correct recognition as a binary dependent variable (1/0) was conducted to analyse change in emotion recognition accuracy with age, by audio condition and emotion. Independent fixed effects were condition, emotion, and age, and the interactions between these variables. Participant and stimuli were entered as random intercepts (see Appendix F for model selection procedure). Due to the binary dependent variable, post-hoc comparisons were analysed in terms of odds ratios. Overall emotion recognition accuracy for each audio condition was then correlated to assess associations between conditions. Finally, multiple regression models were fit with overall recognition accuracy for either instrumental or prosody conditions as the dependent variable, to assess whether accuracy for audio conditions were related independent of age and music training. Independent variables were recognition accuracy for singing stimuli, and accuracy for instrumental or prosody stimuli

(depending on which was the dependent variable), while controlling for age and years of music training.[11]

3.3.3.1.2 Associations with Stimulus Arousal and Valence

To analyse the influence of stimulus valence and arousal levels on emotion perceptions, a series of GLMMs were conducted. For each emotion category, emotion selection (yes/no) was the binary dependent variable. Stimulus arousal and stimulus valence, and their interactions with condition and age, were entered as independent variables. This allowed inferences regarding the relationship between arousal or valence and emotion perceptions while controlling for the other dimension – drawing out their unique effects (e.g., the effect of stimulus arousal on emotion selection irrespective of positive or negative valence). A systematic relationship between arousal/valence and the likelihood of selecting a given emotion would indicate that these properties are associated with the perception of that emotion. Two models were specified for each emotion – one across the whole set of emotions (full model), and one excluding the target emotion (errors only model). This allowed inferences to be drawn regarding how stimulus valence and arousal levels relate to participants' perceptions of each emotion overall, and how they influence erroneous emotion perceptions (i.e., confusion patterns). To address hypothesised relative consistency between conditions and with age for arousal relative to valence, summary tables and figures of results for a) interactions between valence/arousal and condition, and b) interactions between arousal/valence and age, are presented.

[11] A model with singing as the dependent variable was not conducted, given singing's proposed role as a developmental basis for later developing vocal emotion recognition ability (Schubert & McPherson, 2015).

3.3.3.2 Bayesian Analyses.

BFs were included for a) correlation analyses, b) GLMM and regression analyses at the level of fixed effects (including post-hoc comparisons). All Bayesian models were fit with matching fixed and random effects structures to the frequentist models above, using the default priors established in past research (Morey et al., 2015; Oberaur, 2019, 2023). For mixed models, BFs were not included for random effects as they were not specified with default priors, as is required for accurate bayes factors. BFs were not included for the GLMM analyses of the prediction of emotion perceptions by stimulus arousal and valence. This was due to instability in BF estimates (different estimates with each computation), possibly due to model complexity. This was the case even at high numbers of model iterations (50,000). All analyses were conducted in R (R Core Team, 2022). Further details of Bayesian analyses, including prior specifications, as well as full details of statistical procedures, including assumption tests and reporting, can be seen in Appendix E.

3.4 Results

3.4.1 Emotion Recognition

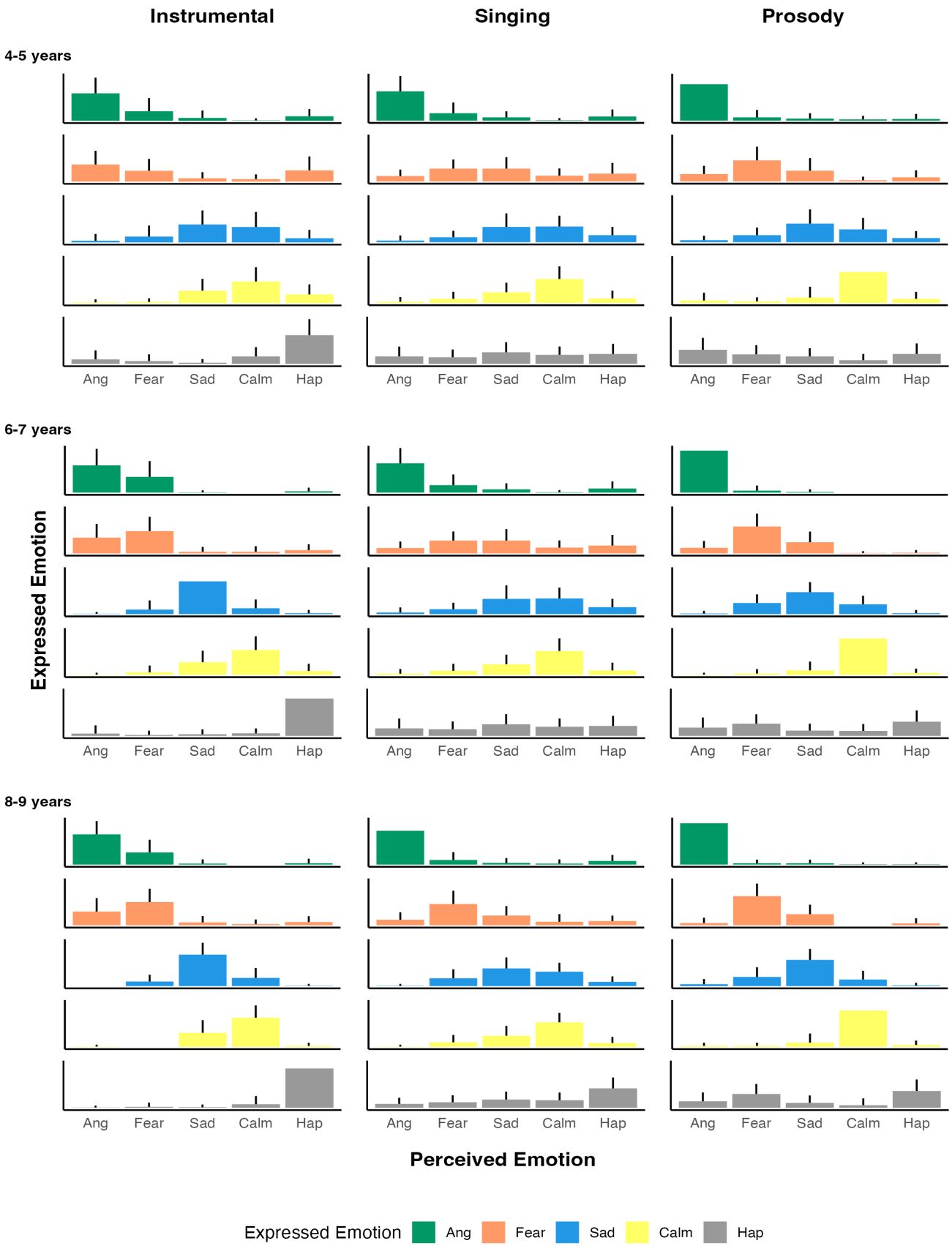
3.4.1.1 Average Raw Accuracy by Condition, Emotion, and Age-Group.

Figure 3.3 outlines mean emotion recognition accuracy and confusion patterns for each condition and age-group.[12] Emotion recognition patterns were most similar between conditions in 4-5-year-olds, with anger and calmness generally high in accuracy within each condition relative to other emotions. This age group displayed relatively diverse confusion patterns, including between emotions similar in arousal and between emotions similar in valence. While some confusion patterns were consistent across conditions, such as sadness and calmness, others were only apparent in certain conditions. For example, fear was most

[12] Tables of raw emotion recognition accuracy scores are in Appendix K.

often erroneously selected as sadness for singing and prosody stimuli, but as anger or happiness for instrumental stimuli. For 6-7-year-olds, these confusion patterns remained but to a lesser extent, while further differences between conditions also emerged (e.g., higher anger accuracy for prosody relative to instrumental music). By 8-9-years, most confusion was between emotions hypothetically similar in arousal. However, fear and sadness were still confused for prosody stimuli, while singing displayed wider confusion patterns than other conditions.

Figure 3.3 - Raw Mean Emotion Recognition Accuracy (%) and Confusion Patterns by condition, emotion, and age-group.



Note. Error bars = standard deviation. The main diagonal represents correct emotion recognition. Ang = anger; Sad = sadness; Hap = happiness. *N* = 95 (4-5 years = 27; 6-7 years = 34; 8-9 years = 34). X-axis = % correct.

To analyse emotion recognition patterns, GLMM analysis was conducted with correct recognition (0 or 1) as the dependent variable.[13] Independent variables were emotion, condition, age, and the interactions between these variables. Stimuli and participant were entered as random intercepts. The random intercept for stimuli significantly improved model fit (c^2 (1) = 181.81, p<.001), as did the intercept for participant (c^2 (1) = 51.68, p<.001).

3.4.1.2 Emotion and Condition Effects.

There was a significant main effect of emotion (c^2 (4) = 41.96, p<.001, BF_{10} = 7.13). The odds of correct recognition were higher for anger compared to all other emotions, with moderate evidence for higher odds compared to calmness (OR = 2.12, *SE* = 0.50, p =.01, BF_{10} = 8.26), and decisive evidence for higher odds compared to all other emotions (ps<.001, BF_{10s} >100). Calmness was also more accurately recognised than sadness (OR = 1.91, *SE* = 0.44, p =.04, BF_{10} = 1.77), happiness (OR = 2.14, *SE* = 0.50, p =.01, BF_{10} = 3.73), and fear (OR = 2.20, *SE* = 0.51, p =.005, BF_{10} = 4.12).

There was a significant main effect of condition (c^2 (2) = 17.39, p<.001, BF_{10} = 2.76). The odds of correct recognition were higher for instrumental music compared to singing (OR = 1.94, *SE* = 0.35, p<.001, BF_{10} = 9.55) and for prosody compared to singing (OR = 2.17, *SE* = 0.39, p<.001, BF_{10} = 27.48), but did not differ between instrumental music and prosody (OR = 1.13, *SE* = 0.21; p =.79, BF_{10} <.001).

There was a significant interaction between condition and emotion (c^2 (1) = 44.56, p<.001, BF_{10} >100). The odds of correct anger recognition were higher for prosody relative to instrumental music stimuli (OR = 6.76, *SE* = 2.91, p <.001, BF_{10} = 8.29). Odds were also

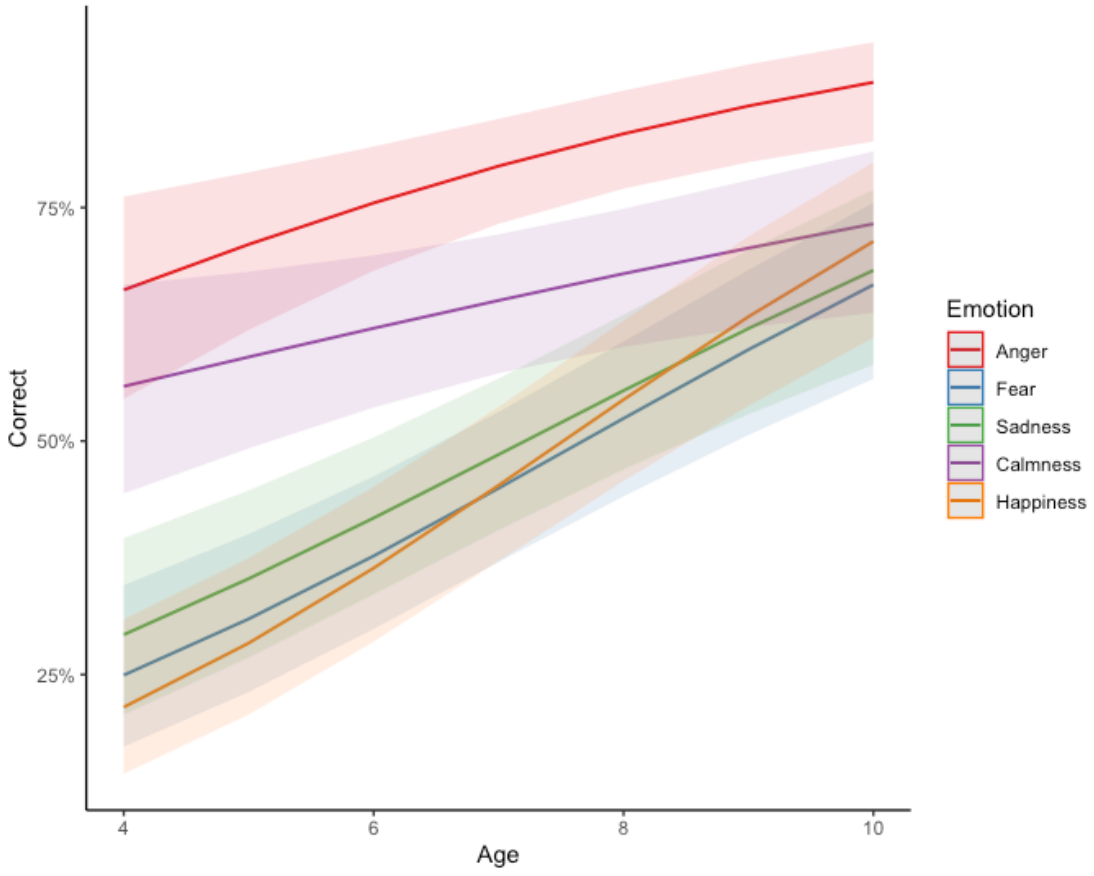
[13] An initial t-test indicated no difference in overall recognition accuracy between males and females (t (91.09) = 1.82, p =.072). Sex was not included in following analyses.

higher than singing, but this did not reach statistical significance (OR = 3.95, *SE* = 1.71, *p* =.09, BF_{10} = 1.39). The odds of correct recognition were higher for instrumental happiness compared to prosody and singing stimuli (*p*s <.001, BF_{10s} >100).

3.4.1.3 Age-Related Effects.

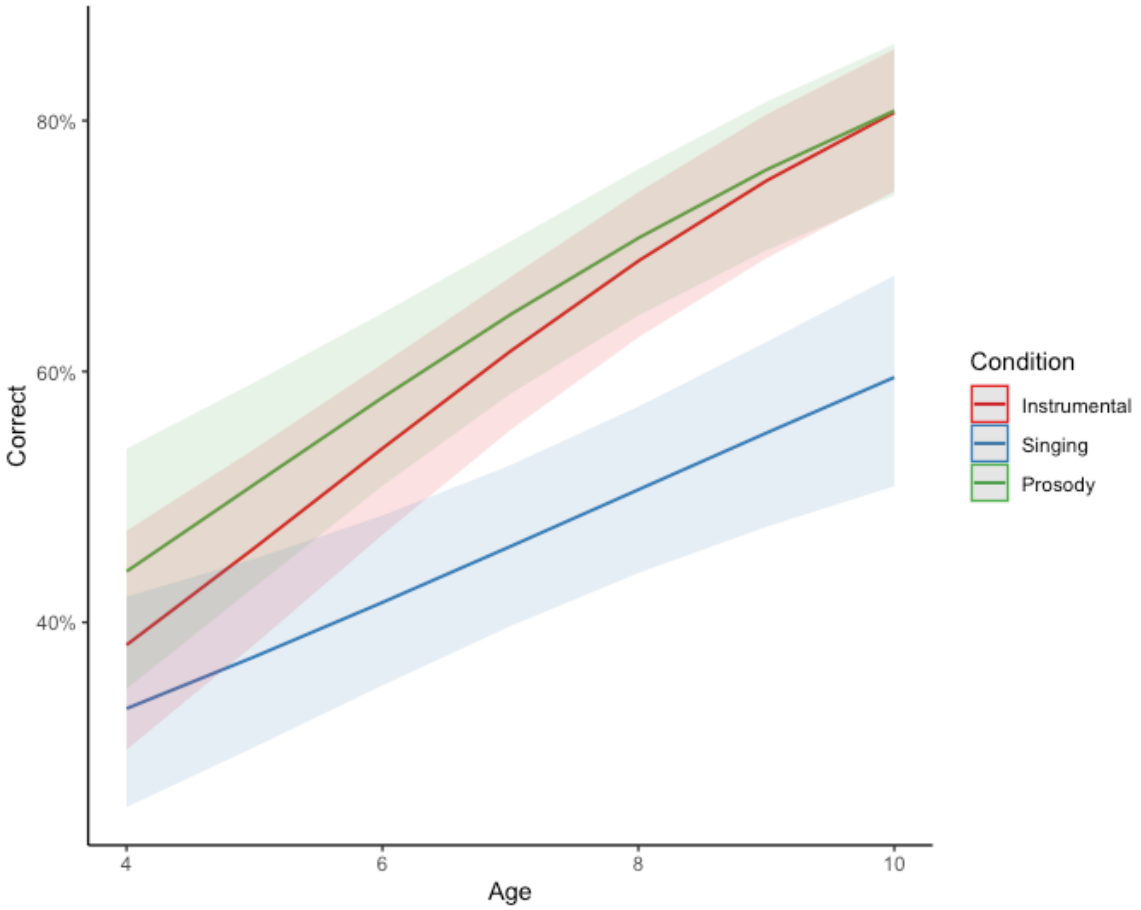
The GLMM found a significant main effect for age, with the odds of correct recognition increasing with age (c^2 (1) = 45.31, *p*<.001, BF_{10} >100). There was also a significant interaction between emotion and age, although the BF indicated only anecdotal evidence for this effect (c^2 (4) = 18.78, *p*<.001, BF_{10} = 1.99). As indicated in Figure 3.4, happiness had the greatest increase in the odds of correct recognition with age, and this was significantly higher than for calmness, which had the flattest trajectory (*z* =3.46, *p* =.005, BF_{10} = 35.43).

Figure 3.4 – Line Graph with Confidence Intervals Showing Marginal Predicted Probability of Emotion Recognition Accuracy by Age and Emotion.



There was also a significant interaction between condition and age, although the BF indicated anecdotal evidence *against* this effect (c^2 (2) = 8.33, p =.02, BF_{10} = 0.34). As shown in Figure 3.5, singing showed a slightly flatter developmental trajectory than instrumental music (z = -2.67, p =.02, BF_{10} = 1.53), while there was evidence against a difference between other conditions (ps >.05, BF_{10s} <.03).

Figure 3.5 – Line Graph with Confidence Intervals Showing Marginal Predicted Probability of Emotion Recognition Accuracy by Age and Condition.



There was no three-way interaction between condition, age, and emotion (c^2 (8) = 12.73, p =.12, BF_{10} = 0.43).

3.4.2 Associations Between Emotion Recognition Accuracy, Age, and Music Training

Table 3.1 outlines correlations between recognition accuracy for each condition, age, and years of music training. There were relatively strong correlations between all accuracy variables and between all accuracy variables and age. There was anecdotal evidence against a correlation between music training and recognition accuracy for singing and prosody. While

the correlation between music training and instrumental recognition accuracy was statistically significant, the BF provided no conclusive evidence for or against this relationship.

Table 3.1 – Correlations Between Recognition Accuracy, Age, and Years of Music Training

Variable	1	2	3	4
1. Accuracy Instrumental	-			
2. Accuracy Singing	.45*** (>100)	-		
3. Accuracy Prosody	.48*** (>100)	.43*** (>100)	-	
4. Age	.57*** (>100)	.42*** (>100)	.46*** (>100)	-
5. Music Training (Years)	.25* (1.01)	.02 (0.32)	.13 (0.76)	-

Note. Coefficient (BF_{10}). Spearman's rho correlations involving music training and age, Pearson's correlations between accuracy variables. $df = 95$. $p<.05$*, $p<.01$**, $p<.001$***.

A hierarchical multiple regression was conducted to assess whether emotion recognition accuracy could be predicted by accuracy in other conditions, independent of age and music training. This was done with prosody recognition and instrumental recognition as the dependent variables, within separate models. For each model, step 1 included age and music training, and accuracy scores were added at step 2.

For prosody, at step 1, age and music training explained 21.1% of variance in emotion recognition accuracy ($F(2, 79) = 11.86$, $p < .001$, $BF_{10} > 100$). Age was a significant predictor ($t(79) = 4.70$, $p < .001$, $BF_{10} > 100$), while music training was not ($t(79) = 0.67$, $p = .51$, $BF_{10} = .27$). At step 2, the model explained 32.7% of variance in prosody recognition scores ($F(4, 77) = 10.84$, $p < .001$, $BF_{10} > 100$). While holding all other predictors constant, the only

significant predictor of prosody emotion recognition accuracy was singing accuracy, and the BF provided moderate evidence for this effect ($t(77) = 2.85$, $p = .005$, $BF_{10} = 9.23$). There was no evidence for or against an effect for age at this step ($t(77) = 1.93$, $p = .06$, $BF_{10} = 1.41$). There was anecdotal evidence against an effect for instrumental recognition accuracy ($t(77) = 1.54$, $p = .13$, $BF_{10} = .77$), and moderate evidence against one for music training ($t(77) = 0.63$, $p = .53$, $BF_{10} = 0.31$.

For instrumental music, at step 1, age and music training explained 34.6% of the variance in emotion recognition scores ($F(2, 79) = 22.4$, $p < .001$, $BF_{10} > 100$). Age was a significant predictor in this model ($t(79) = 6.25$, $p < .001$, $BF_{10} > 100$), while music training was not ($t(79) = 1.59$, $p = .18$, $BF_{10} = 0.57$). At step 2, the model explained 43.1% of variance ($F(4, 77) = 16.35$, $p < .001$, $BF_{10} > 100$). While holding all other variables constant, age remained a significant predictor of instrumental accuracy ($t(77) = 3.78$, $p < .001$, $BF_{10} > 100$), and singing recognition accuracy was also a significant predictor ($t(77) = 2.56$, $p = .01$, $BF_{10} = 4.29$). There was anecdotal evidence against an effect for music training and prosody recognition accuracy ($ps > .05$, $BF_{10s} < 1$).

3.4.3 Associations with Stimulus-Level Valence and Arousal

To analyse associations between stimulus valence and arousal levels and emotion perceptions, a series of GLMMs were conducted. For each emotion, the influence of stimulus arousal and stimulus valence on the likelihood of selecting the given emotion, and whether this varied by age or condition, was assessed. Each model controlled for the effect of the other dimension, to assess the unique effects of arousal and valence. Two models were specified for each emotion – one across the whole set of emotions (full model), and one excluding the target emotion (errors only models). It was hypothesised that arousal would predict emotion perceptions relatively consistently across conditions and show limited

increase with age. Valence was expected to display more variation between conditions, and more age-related increase, compared to arousal.

3.4.3.1 The Effect of Stimulus Arousal and Valence by Condition.

There were significant main effects for arousal and valence on the odds of emotion selection for each emotion category, for both full and errors only models (*p*s <.01). Table 3.2 highlights the models that had significant interactions between arousal or valence and condition. There were significant interactions between condition and stimulus arousal for all full models, and all but happiness for errors only models. For valence, there was no significant interaction with condition for anger when all target emotions were included, and no interaction for sadness when considering only error patterns.

Table 3.2 – Interactions Between Stimulus Arousal and Condition, and Stimulus Valence and Condition, on the Odds of Emotion Selection.

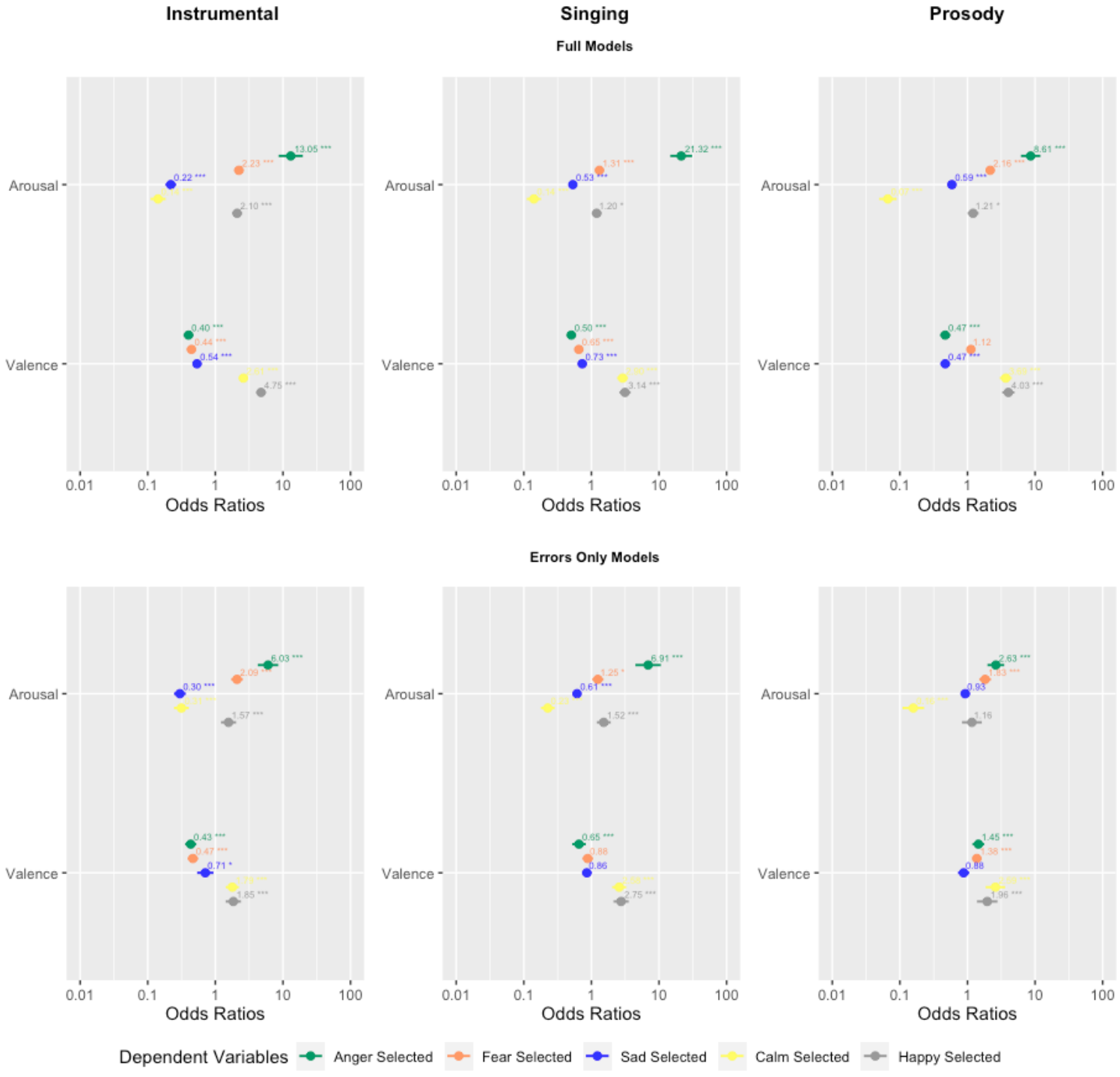
Emotion	**Stimulus Arousal*Condition**		**Stimulus Valence*Condition**	
	Full	**Errors**	**Full**	**Errors**
Anger	+***	+***	/	+***
Fear	+***	+***	+***	+***
Sadness	+***	+***	+***	/
Calmness	+***	+***	+*	+***
Happiness	+***	/	+***	+***

Note. *df* = 2. *p*<.05*, *p*<.01**, *p*<.001***. + = significant interaction with condition; / = no significant interaction with condition. Interaction effect = Likelihood Ratio Chi Square Statistic.

Figure 3.6 explores these interactions further. The condition specific patterns for arousal are apparent, but these relate to the strength, rather than direction, of effect, and arousal effects were generally stronger than valence effects. The main exception to this was

happiness, which showed a stronger relationship with stimulus valence compared to arousal across models. Most between-condition differences in the effect of valence related to strength of effect, but valence was positively related to anger and fear selection in prosody error models (more positive valence = more erroneous anger and fear selections). This pattern was not apparent for other audio conditions.

Figure 3.6 – Effects of Arousal and Valence on the Odds of Emotion Selection, for Each Condition.



Note. $p<.05$*, $p<.01$**, $p<.001$***. Each box represents an individual generalised linear mixed model, with either instrumental music, singing, or prosody conditions held as the reference category. Age held at mean for all models. Full models = including target emotion; Errors only models = excluding target emotion.

3.4.3.2 Age Interactions.

There were also significant positive interactions between stimulus arousal/valence and age. Table 3.3 outlines these interactions. For full models, there was an increase in the effect of stimulus valence on the odds of emotion selection with age, for all emotions. This was not the case for arousal, with anger and calmness showing no interaction with age. Within error patterns, anger and calmness showed an increase in the effect of stimulus arousal on emotion selection with age, while only anger showed this pattern for stimulus valence.

Table 3.3 – Interactions Between Stimulus Arousal and Age, and Stimulus Valence and Age, on the Odds of Emotion Selection.

Emotion	**Stimulus Arousal*Age**		**Stimulus Valence*Age**	
	Full	**Errors**	**Full**	**Errors**
Anger	/	+*	+***	+*
Fear	+*	/	+*	/
Sadness	+***	/	+***	/
Calmness	/	+***	+*	/
Happiness	+**	/	+***	/

Note. df = 1. $p<.05$*, $p<.01$**, $p<.001$***. + = significant positive interaction with age; / = no significant interaction with age. Interaction effect = Likelihood Ratio Chi Square Statistic (c^2).

3.5 Discussion

The present Chapter explored similarities and differences in emotion recognition patterns for instrumental, singing, and prosody stimuli in TD 4-9-year-old children. Stimulus arousal and valence levels were analysed in relation to perceptual patterns, and any differences in these associations based on audio condition and age were examined. Findings indicated similarities in emotion recognition for audio conditions, with correlations between

accuracy for all conditions, and similar overall accuracy and rates of development for instrumental music and prosody. While singing was less accurately recognised overall, it significantly predicted accuracy in both other audio conditions independent of age and music training, suggesting developmental significance of emotion understanding of singing stimuli (Schubert & McPherson, 2015). Findings also offered some support for the hypothesised broad-to-differentiated model of emotion recognition development (Widen, 2013) that centralises comprehension of expressed arousal (Holz et al., 2021). However, between-condition differences in relation to how both arousal and valence relates to emotion perceptions were apparent, while there were also differences between conditions in recognition accuracy for certain emotions. The implications of these between-condition similarities and differences are discussed in relation to a proposed cross-condition model of audio emotion recognition development (Heaton & Allgood, 2015).

In line with hypotheses, there were similarities between audio conditions in how emotion recognition accuracy developed. Accuracy for all three conditions were positively correlated with each other and increased with age, while similarities in the rate of change were particularly pronounced between instrumental and prosody stimuli. This aligns with past research indicating parallel typical development of instrumental and prosodic emotion recognition accuracy (Heaton & Allgood, 2015; Vidas et al., 2018). Findings also support the presence of an important developmental period for audio emotion recognition development between 4 and 9 years, in line with past vocal prosody and instrumental music research (Chronaki et al., 2015b; Grosbras et al., 2018; Heaton & Allgood, 2015). This also aligns with facial expression research, where the period before 10-years appears particularly important for emotion recognition development (Chronaki et al., 2015b). Further, although there were differences between conditions in overall accuracy for certain emotions, there was no evidence for a difference in the rate of change for specific emotions between conditions.

This suggests that emotions reached more 'adult-like' levels at a similar rate across audio conditions. Accordingly, the present findings align with the possibility of a cross-condition model of audio emotion recognition development (Heaton & Allgood, 2015). This could be linked to functional similarities in how musical and vocal emotions are expressed and processed (Escoffier et al., 2013; Frühholz et al., 2016; Juslin & Laukka, 2003; Vigl et al., 2024), perhaps due to a shared evolutionary basis in primitive forms of emotion expression (Brown, 2017; Clark et al., 2015).

Emotion recognition findings for singing stimuli aligned less closely with other audio conditions, but also suggested a possible unique role for singing stimuli within audio emotion recognition development. Singing was less accurately recognised than both instrumental and prosody stimuli overall, while accuracy was most closely aligned with other audio conditions earlier in development. This aligns with past research indicating similarities in emotion recognition accuracy between singing and instrumental stimuli in young children (Franco et al., 2017), but relative difficulty recognising sung emotions in older children and adults (Chapter 2; Dolgin & Anderson, 1990; Livingstone & Russo, 2018). However, despite children's relative difficult with emotion recognition for singing stimuli overall, it significantly predicted accuracy for both instrumental and prosody conditions, even when controlling for level music training, age, and accuracy for the other respective audio condition. This could be linked to singing's specific role within development. Infant-directed song often involves exaggerated expression of acoustic features (Nakata & Trehub, 2011), while past research has suggested that the early ability to draw meaning from early 'song-like' mother-child interactions may form the basis of developing emotion understanding across audio conditions (Boone & Cunningham, 1998; Schubert & McPherson, 2015; Vanden Bosch der Nederlanden et al., 2022). Seemingly, although less salient in terms of discrete emotions in general, the ability to comprehend emotional information in singing is an

important developmental skill. Future longitudinal research exploring links between recognition accuracy for singing and other audio conditions across development could be illuminating in this regard.

The present chapter found some condition-specific accuracy patterns for certain emotions. In line with past research with children, prosodic anger was salient from early in development (Nelson & Russell, 2011; Sauter et al., 2013; van Zonneveld et al., 2019) and easier to recognise than musical anger (Vidas et al., 2018). Conversely, happiness was easier to recognise in instrumental music compared to prosody, aligning with past research suggesting early salience of instrumental happiness (Stacho et al., 2013), but relative difficulties for prosody that persist into adulthood (Chapter 2; Nelson & Russell, 2011). However, sadness was relatively difficult to recognise overall in the present chapter and seemed particularly challenging for prosody and singing stimuli. This misaligns with some past research indicating that sadness is one of the easiest prosodic emotions to recognise from early in development (e.g., Grosbras et al., 2018; Nelson & Russell, 2011). However, the present chapter's inclusion of calmness may have contributed to this relative difficulty, as most past studies included sad as the only hypothetically low arousal emotion (Chronaki et al., 2015b; Grosbras et al., 2018; Nelson & Russell, 2011). Indeed, confusion between low arousal emotions sadness and calmness was common across audio conditions, particularly in younger children. This may suggest that arousal comprehension is key to early emotion recognition development, in line with theoretical and empirical research highlighting the salience of arousal properties in vocal and musical emotion expressions and perceptions (Bänziger et al., 2015; Cespedes-Guevara & Eerola, 2018; Holz et al., 2021; Wenninger et al., 2013).

Associations between stimulus arousal/valence and emotion perceptions in the present chapter add clarity regarding the role of these affective features during audio emotion

recognition development. As hypothesised, stimulus arousal was relatively strongly related to emotion perceptions across audio conditions, such that happiness, anger, and fear selections were more common as stimulus arousal increased, and sadness and calmness selections were more common as stimulus arousal decreased. Although the effect of stimulus arousal did vary between conditions for most emotions, this was due to differences in the strength of effect, which was generally strong across conditions. This indicates a prominent role for stimulus arousal within audio emotion perceptions, in line with past research (Cespedes-Guevara & Eerola, 2018; Holz et al., 2021; Wenninger et al., 2013). Perceptions of anger and calmness were most strongly predicted by stimulus arousal, while the effect of arousal did not increase with age for these emotions. Within each condition, these emotions were also the easiest for younger children to recognise and showed relatively flat emotion recognition developmental trajectories. Across audio conditions, these emotions differ greatly in perceived arousal (Chapter 2). Indeed, some past research has indicated that these emotions can be accurately represented as opposing ends of a continuous plane of expressed arousal (Woodward et al., 2021). Accordingly, the specific relationships between stimulus arousal and anger and calmness perceptions supports the possibility that arousal-based understanding is particularly important to audio emotion recognition early in development (Kragness et al., 2021; Nelson & Russell, 2011; Zupan, 2015) – differing from the stronger role of valence within early facial emotion recognition development (Widen, 2013).

However, contrary to hypotheses, there were interactions between stimulus arousal and age for most emotions. This suggests that the relationship between stimulus arousal and the perception of some emotions increases with age. This aligns with facial expression research, where although valence is the most prominent affective feature affecting young children's emotion perceptions, it's role within emotion perception development continues to increase with age (Woodward et al., 2022). Accordingly, although arousal may be particularly

prominent within early audio emotion perceptions, arousal-based understanding and its role in substantiating emotion perceptions may continue to develop throughout childhood.

Hypothesised age-related increase in the relationship between stimulus valence and emotion perceptions was supported overall. As age increased, happiness and calmness selection became more common for stimuli more positive in valence, while anger, fear, and sadness selection became more common for stimuli more negative in valence. This developing sensitivity can partially explain the reduction in some between-valence emotion recognition confusion across conditions, such as selection of fear as happiness, and happiness as anger. Given early sensitivity to arousal-related acoustic features such as loudness and tempo in musical stimuli (Dalla-Bella et al., 2001; Kragness et al., 2021), and strong relationships between these features and arousal in vocal stimuli (Scherer et al., 2015), findings align with proposed early developing, automatic arousal-based perceptions, with later-developing valence understanding and emotion concept knowledge allow increasingly refined emotion recognition (Nelson & Russell, 2011; Woodward et al., 2021; Woodward et al., 2022).

However, between-condition differences in the relationship between stimulus valence and some emotions pointed to complexity regarding how this valence understanding may develop across audio conditions. As with arousal, most between-condition differences in the association between valence and emotion perceptions related to differences in strength of effect. However, there were also cases of a more pronounced difference in the direction of this valence effect between conditions. For example, erroneous anger and fear perceptions for prosody stimuli were related to increased (more positive) valence. This was reflected in their confusion with expressed happiness, while this pattern was not apparent for instrumental or singing stimuli. This aligns with general biases towards negative emotions in past prosody research with children, which may be indicative of limited valence-based understanding for

vocal emotions, particularly early in development (Vaish et al., 2008; Nelson & Russell, 2011). Indeed, Holz et al (2021) suggest that valence may act as an 'affective semantic' indicator that allows refinement of more automatic arousal-based perceptions. While for musical stimuli this process of refinement can draw on salient expressive cues such as musical mode (Gomez & Daunser, 2007), valence in prosody may be more of a conceptual cue (based on knowledge of positive/negative emotions) given its limited links to acoustic properties (Chapter 2; Sauter et al., 2010; Weninger et al., 2013). Collectively, this may lead to differences in the role of valence during emotion recognition for musical and vocal stimuli. Indeed, sensitivity to musical mode is thought to develop around 6-8-years (Dalla-Bella et al., 2001), and this likely related to reductions in confusion between happiness and anger/fear between younger and older age-groups for instrumental (and to a lesser extent singing) stimuli, while confusion between these emotions for prosody stimuli persisted with age.

It is important to note that cross-condition emotion recognition development cannot be reduced absolutely to understanding of these fundamental affective dimensions. Indeed, as discussed, children's conceptual knowledge of emotions plays an integral role in how variations in these affective dimensions are interpreted as emotions – facilitating increasingly differentiated emotion recognition abilities (Woodward et al., 2022). Further, children form functional associations between emotion expressions and certain behavioural outcomes, and begin to distinguish them on perceptual properties, before they develop ability to distinguish them on arousal or valence properties (Ruba et al., 2018; Ruba & Repacholi, 2020). This suggests that links between expressed features and emotion perceptions, independent of understanding of affective dimensions, are possible. Indeed, certain acoustic features can communicate vocal anger independent of perceived arousal (Bänziger et al., 2015; Giordano et al., 2021). This can explain the higher relative accuracy for prosodic anger in the present chapter. Given the adaptive significance of quickly comprehending and responding to

expressed anger (Buss, 2005), it may be that early functional associations between perceptual features of anger and certain outcomes facilitates early comprehension of prosodic anger. This anger comprehension may develop independent of the proposed broad-to-differentiated developmental emotion recognition process, leading to salience of anger in vocal prosody, specifically. Given the unique social function involved in this link between expressed vocal anger and emotion recognition, it is likely that a similar processing approach is less prominent for musical stimuli. Instead, emotion recognition for musical anger may follow the hypothesised broad-to-differentiated pathway - producing the music-specific confusion between anger and fear in the present chapter given their similarity in both arousal and valence (Chapter 2). Indeed, stimulus arousal and valence were each more strongly related to anger and fear perceptions for instrumental music, relative to vocal prosody stimuli. Seemingly, while consideration of arousal and valence dimensions can aid inference regarding audio emotion recognition patterns, the relative role of these affective dimensions may differ in subtle ways between conditions.

3.5.1 Limitations and Future Directions

The present chapter had a range of limitations. First, direct comparisons between stimulus valence and stimulus arousal in terms of their predictive strength on emotion perceptions (i.e., claims that arousal was stronger than valence in its effect) should be treated with caution due to the limited number of stimuli. A larger set of stimuli would lead to greater levels of variance for each affective dimension and facilitate more accurate inferences regarding the relative sensitivity of emotion perceptions to variation in each dimension. Similarly, consideration of a wider set of emotion categories would allow firmer inferences regarding the role of stimulus arousal/valence within error patterns, due to an increase in the number of emotions closely matched in arousal/valence. Although this may be challenging within a cross-condition paradigm given the limited number of validated musical stimuli sets

including a wider number of emotions, some recent research has provided evidence that emotions such as pride and love can be accurately recognised in music (Micallef Grimaud & Eerola, 2021; Vidas et al., 2018). Analyses of associations between stimulus arousal/valence and emotion perceptions also do not provide evidence regarding how increasing sensitivity to these dimensions may relate to recognition *accuracy*. Indeed, in some cases, higher sensitivity to valence with age may have led to more confusion between emotions similar in both arousal and valence (e.g., between anger and fear for instrumental music). Thus, developing sensitivity to arousal and valence is just one aspect of audio emotion recognition development, while other factors such as emotion concept knowledge (Widen, 2013) and direct links between expressive acoustic features and emotion recognition (Bänziger et al., 2015) also play integral roles, particularly when differentiating emotions similar in both arousal and valence.

The present findings are also limited by the methodological approach to assessing emotion recognition accuracy. While forced-choice approaches represent an important paradigm to measure a key developmental skill, they do not reveal information regarding children's understanding of emotion expressions beyond pre-determined emotion categories (Barrett, 2017). For example, it may be that children have some arousal-based understanding of audio emotion expressions independent of their alignment with an emotion category, in line with similar evidence for facial expressions (Woodward et al., 2022). Free-sorting approaches where children group stimuli based on how similar they perceive them to be (e.g., Woodward et al., 2022), that require no assumptions regarding expressed emotion categories, could be illuminating regarding the extent to which children rely on arousal, valence, and emotion category knowledge to perceive and represent emotion stimuli. Although challenging with audio stimuli, certain technological methodological advances (e.g., see Donhauser & Klein, 2022) are making this approach feasible for audio stimuli. Such an approach would be

particularly useful for directly contrasting the strength of the influence of valence, arousal and emotion concept knowledge in how children represent and understand emotion stimuli - potentially corroborating the proposed relative prominence of arousal in audio relative to visual stimuli.

3.5.2 Conclusion

Overall, findings offer some support for a condition-general model of audio emotion recognition development, with similarities in overall emotion recognition accuracy and developmental trajectories for instrumental and prosodic stimuli. Incorporation of singing into such a model appears well-founded, with emotion recognition for singing stimuli predicting accuracy in other audio conditions, aligning with the proposed important developmental role of understanding expressed affect in singing (Corbel et al., 2016; Shenfield et al., 2003). Findings also offer some support for the proposal that emotion recognition development follows a broad-to-differentiated trajectory, involving a developing ability to interpret arousal and valence dimensions alongside improving conceptual understanding of emotions (Widen, 2013). Findings also suggest that understanding of arousal may be particularly important for audio emotion recognition development, in line with past research with adults (Chapter 2; Holz et al., 2021; Weninger et al., 2013). However, music-specific development of sensitivity to prominent valence indicators such as musical mode (Dalla-Bella et al., 2001), and possible direct links between some vocal cues and perceptions of emotions such as anger (Bänziger et al., 2015; Ruba et al., 2018), may give rise to some condition-specific recognition accuracy and confusion patterns. Further, given the importance of developing conceptual understanding of emotions to the broad-to-differentiated model of emotion recognition development (Widen, 2013), understanding of how this element relates to audio emotion recognition development appears an important avenue for future research.

4. Individual Differences in Children's Emotion Recognition Instrumental Music, Singing, and Vocal Prosody – The Role of Emotion Language Comprehension

4.1 Relationship with Previous Chapters

Chapter 2 explored adults' perceptions of valence and arousal and emotion recognition accuracy for instrumental music, singing, and vocal prosody stimuli. A set of acoustic features were also extracted from the stimuli and analysed as possible partial explanations for patterns of perception. Chapter 3 extended this research to typically developing 4-9-year-old children to explore similarities and differences in developmental trajectories for each stimulus type. Stimulus-level affective features - normative adult-rated levels of valence and arousal (from study 1) - were also analysed in relation to emotion recognition patterns.

Findings revealed positive associations between overall instrumental, singing, and prosody emotion recognition accuracy, while children's accuracy for instrumental and prosody conditions developed at a similar rate. Adults' perceptions of valence and arousal were also similar between conditions. However, for both adults and TD children, there were condition-specific patterns for certain emotions such as anger (more salient in prosody) and happiness (more salient in instrumental music). While these patterns could be attributed, in part, to stimulus-level features, there remained unanswered questions, particularly in terms of some of the mechanisms at the cognitive level that may underpin similarities and differences between audio conditions. Understanding the role of these cognitive processes could be key to further specifying a cross-condition developmental model of audio emotion recognition, while providing a basis for exploring individual differences in audio emotion recognition development – key to any music-based intervention approaches. Accordingly, the present

chapter focuses on emotion language comprehension as one general mechanism of musical and vocal emotion recognition development.

4.2 Introduction

4.2.1 Musical and Vocal Emotion Recognition Development – Individual Differences and Possible Mechanisms

Music is closely related to vocal prosody via their parallel capacity to elicit and convey emotion (Juslin & Sloboda, 2011), linked to a theorised functional evolutionary basis to musical and vocal emotion emotions (Juslin, 2018; Scherer, 1995; 2017; Brown, 2017). This translates to similarities in emotion recognition development. Indeed, in TD children, emotion recognition for instrumental music, singing, and vocal prosody are correlated (Chapter 3; Franco et al., 2017), while emotion recognition accuracy for instrumental and prosody stimuli appears to develop in parallel (Chapter 3; Heaton & Allgood, 2015; Vidas et al., 2018). For Heaton & Allgood (2015, p. 402), this evidence points to a 'cross-condition model of auditory emotion recognition'. However, past developmental research on links between musical and vocal emotion recognition has predominantly been with TD samples. Further, beyond reference to similarities in acoustic features (Paquette et al., 2018) and consideration of levels of music training (Chapter 3; Vidas et al., 2020), developmental mechanisms remain unexplored. It is thus unclear whether emotion recognition accuracy for these audio conditions tends to covary within samples of children with more varied levels of socioemotional difficulties, and how such covariation could be explained.

Individual differences may arise within various stages underpinning musical and vocal emotion recognition. This includes those supporting the low-level perception of acoustic information, and those supporting the higher-level cognitive interpretation of emotional meaning (Bestelmeyer et al., 2014; see section 1.6.3 for a discussion). Although

some music-specific (Whitehead & Armony, 2018) and voice-specific (Belin et al., 2002) neural processing mechanisms have been identified (Whitehead & Armony, 2018), extensive overlap in neurophysiological responses have been demonstrated during *emotion* processing of musical and vocal stimuli (Escoffier et al., 2013; Frühholz et al., 2016; Paquette et al., 2018). Importantly, in alignment with Bestelmeyer et al.'s (2014) model, overlap is apparent both during acoustic integration (Peretz et al., 2015) and the conceptual representation of emotion stimuli (Schirmer et al., 2012).

Past research on vocal and musical emotion recognition has mainly focused on individual differences in the low-level acoustic-perceptual aspect of this model. Given evidence for similar patterns of acoustic features such as loudness and tempo/speech rate for some musical and vocal emotions (Juslin & Laukka, 2003), individual differences in sensitivities to these acoustic features have been proposed as central to both musical and vocal prosody emotion recognition. Indeed, evidence suggests that musical aptitude – defined in this case as the ability to perceive differences in acoustic qualities such as pitch, loudness, tempo, etc – relates to vocal emotion recognition ability, independent of level of formal music training (Jansen et al., 2023). Further, this relationship is mediated via an improved ability to detect differences in the prosodic aspects of speech (Vigl et al., 2024). Although unexplored in children, this indicates that similar emotion recognition patterns for music and prosody may be explainable via condition-general low-level acoustic processing mechanisms that operate independent of formal music training (Escoffier et al., 2013; Proverbio et al., 2020). However, despite similar overall typical developmental trajectories for music and prosody, past research indicates some condition-specific patterns of recognition accuracy for certain emotions, while some associations between emotion perceptions and acoustic features, such as pitch, also differ (Chapter 2; Laukka et al., 2013a; Laukka et al., 2013b; Vidas et al., 2018). This suggests that this acoustic-perceptual mechanisms offer an incomplete explanation of

similarities and differences in emotion recognition between conditions, and possible individual differences in the development of this ability.

This brings attention to the stage of cognitive interpretation within Bestelmeyer et al.'s (2014) multi-stage model. Indeed, some evidence suggests that emotion recognition for vocal and musical stimuli relies more strongly on general emotional abilities such as emotional intelligence, relative to low-level perceptual ones (Trimmer & Cuddy, 2008) – a finding reflected in neuroimaging research (Escoffier et al., 2013). Further, individuals with congenital amusia (impaired music processing despite normal peripheral auditory processing, cognitive functioning, and music exposure) are impaired in not only their vocal emotion recognition accuracy (Thompson et al., 2012), but also their facial emotion recognition accuracy (Lima et al., 2016). This underlines general socio-emotional processing mechanism common to music and other forms of emotion expression that aligns with the more cognitive process of emotional meaning interpretation (Bestelmeyer et al., 2014). While analogous developmental research is lacking, findings indicate a need to consider these higher-level mechanisms in relation to vocal and musical emotion recognition development. Increasingly, language development is being placed at the core of these developing cognitive aspects of emotion recognition.

4.2.2 Language as a Condition-General Mechanism of Audio Emotion Recognition Development

4.2.2.1 Language and Emotion Recognition Development.

Much research points to a causative link between language and a range of aspects of emotion development (Cole et al., 2010; Harris et al., 2005; Pons et al., 2003). For example, receptive language and literacy ability predict children's facial emotion recognition accuracy, independent of age and a range of other neurocognitive abilities such as attention, theory of

mind, and sensorimotor functions (Beck et al., 2012; Rosenqvist et al., 2014). This has been attributed to language's role in scaffolding the development of emotion knowledge – shaping the way in which emotions are perceived and understood (Shablack & Lindquist, 2019). These findings support an integral role for language in typical facial emotion recognition development.

Research examining associations between language and vocal emotion recognition are limited, as are those considering a wider range of participants beyond typical development. However, findings point to an association analogous to that with facial expressions. Indeed, Griffiths et al. (2020) investigated the longitudinal relationship between children's language abilities at age 5-6 – assessed via a composite measure incorporating receptive, expressive, vocabulary, and narrative components – and their facial and vocal emotion recognition accuracy at age 11-12. They also collected data from a sample of children with Developmental Language Disorder (DLD), but no other diagnosed neurodevelopmental condition or intellectual disability. DLD is not associated with a primary difficulty with socioemotional processing, isolating language-specific effects on emotion recognition accuracy. They found that language ability significantly predicted later emotion recognition accuracy for both facial and vocal stimuli, over and above non-verbal IQ. Further, in line with past research (Boucher, 2000; Taylor et al., 2015), they found that children with DLD had a significant difficulty recognising both facial and vocal emotions, relative to TD children. Collectively, these findings suggest a specific and unique relationship between language and emotion recognition development that extends to vocal stimuli. There is limited research exploring similar associations between language and musical emotion recognition development, but one study did highlight an association between general verbal ability and musical emotion recognition accuracy in 3-6-year-old children (Franco et al., 2017). Considering this finding, as well as similarities in developmental emotion recognition

patterns with vocal prosody (Chapter 3; Heaton & Allgood, 2015; Vidas et al., 2020), linked to domain-general socio-emotional processing mechanisms (Escoffier et al., 2013; Lima et al., 2016), it is plausible to hypothesise shared developmental mechanisms between conditions, too. Further, beyond these links to vocal prosody, research points to a fundamental developmental link between language and music.

4.2.2.2 Language and Music.

Much research points to an inextricable link between the evolution of language and music, with overlap between the structural, cognitive, neurophysiological, and perceptual components of music and language (Fedorenko et al., 2009; Jentschke, 2016; Patel, 2010). This is reflected in developmental patterns for musical and linguistic abilities. Indeed, early sensitivity to various musical components predict language acquisition, while the richness of the home musical environment – including regularity of infant directed song - predicts later language development (Pino et al., 2023; Politimou et al., 2018; Papadamitrou et al., 2021; Franco et al., 2022). Collectively, these findings indicate possible primary links between musical development and language beyond secondary links to vocal prosody. Although untested, it is plausible to suggest that these links may extend to emotional aspects of both language and music, given inherent emotional aspects of early music-like child-caregiver interactions (Trehub & Schellenberg, 1995; Trehub & Trainor, 1998). Indeed, comprehension of emotional components of singing may form the basis of later development of emotion understanding for both instrumental music and vocal prosody (Chapter 3; Pino et al., 2023; Schubert & McPherson, 2015). Considered alongside condition-general socio-emotional processing mechanisms for music and voice (Escoffier et al., 2013; Frühholz et al., 2016; Paquette et al., 2018; Lima et al., 2016), evidence indicates that musical emotion recognition development may be associated with language development in a similar fashion to facial and vocal conditions. However, the mechanisms that underpin links between language and

emotion recognition are less clear (Streubel et al., 2020). Recent research has suggested that the mediating role of developing *emotion-specific* language comprehension may be key to this mechanism.

4.2.2.3 Domain-General or Domain-Specific Language?

Central to the evidence regarding a role for emotion-specific language in emotion recognition development is emotion concept development (Shablack & Lindquist, 2019). Specifically, it is claimed that language-mediated emotion concept development allows children to organise knowledge of different emotions, including their possible causes, consequences, forms of expression, and affective correlates (Barrett, 2006, 2017; Lindquist, 2017). This developing conceptual understanding facilitates a maturation in a range of emotional skills (Nenchevca et al., 2023), including the shift in emotion recognition from a broader dimensional understanding based on arousal/valence, to the ability to categorise a wider range of emotions (Widen & Russell, 2008; Woodward et al., 2021).

Developmental studies examining emotion-specific language in relation to emotion recognition are limited. However, concept development appears key to development in other domains, as well as to typical emotion development more generally. For example, research has highlighted positive associations between domain-specific language, concept knowledge, and proficiency in abilities including theory of mind (Baetsch & Wellman, 1995; de Villiers & de Villiers, 2003; Grazzani & Ornaghi, 2012) and mathematics (Dehaene et al., 1999; Pica et al., 2004). In relation to emotion development, individual differences in TD 7-10-year-olds' emotion vocabulary have been shown to predict overall emotion understanding, independent of age, sex, and general verbal ability (Ornaghi and Grazzani, 2013). This association also appears to extend to facial emotion recognition, specifically (Beck et al., 2012). Crucially, the association between emotion-specific language (as measured by overall emotion vocabulary size) and facial emotion recognition accuracy held when controlling for general verbal ability,

in TD 4-9-year-old children (Streubel et al., 2020). In fact, when considered together, the previously significant effect of general verbal ability was no longer significant, suggesting a specific association between *emotion* language and emotion recognition accuracy (Streubel et al., 2020). This aligns with predictions regarding the role of emotion language as a scaffold for maturing emotion perceptions via concept development (Shablack & Lindquist, 2019). However, the extent to which language-mediated concepts constitute, or interact with, emotion perceptions, remains disputed. Indeed, some research indicates a fundamental role for brain regions involved in higher-level, language-based processing during the perception of low-level features of emotion stimuli. This would suggest interaction between the low-level acoustic integration and higher-level cognitive interpretation stages of Bestelmeyer et al.'s (2014) model. Irrespective of the nature of this interaction, emotion language comprehension appears key to typical emotional development, including facial emotion recognition. However, due to a narrow focus on typical development within past research, it is currently unclear whether this proposed mechanism is robust to higher levels of individual differences, and applicable for children with socio-emotional difficulties.

A further gap in the research base relates to consideration of the relationship between emotion-specific language and audio emotion recognition development. To our knowledge, the only study to consider this relationship involved instrumental music and found that TD 5-6-year-old children's overall instrumental emotion recognition accuracy for calm, fearful, and sad stimuli was significantly predicted by individual differences in their emotion-specific verbal fluency, over and above their general verbal fluency (Plate et al., 2022). This aligns with research involving facial expressions and suggests a possible domain-general role for emotion language comprehension in emotion recognition development. Given the proposed generality of socio-emotional mechanisms underpinning musical and vocal emotion recognition (Escoffier et al., 2013; Lima et al., 2016; Paquette et al., 2018; Proverbio et al.,

2020), plus analogous links between general language and facial, vocal, and musical emotion recognition (Franco et al., 2017; Griffiths et al., 2020), it can be hypothesised that associations with emotion-specific language would also extend to vocal stimuli (both prosody and singing). Indeed, Cespedes-Guevara and Eerola (2018) theorise that these language-mediated conceptual processes sit at the centre of similarities and differences in emotion recognition patterns for music and voice.

Past research has adopted various methods to measure the development of emotion language comprehension in children. These include ratings by parents/teachers regarding the number of words a child understands (e.g., Baron-Cohen et al., 2010), testing fluency by asking children to label as many emotion words as they can (e.g., Beck et al., 2012; Sturrock & Freed, 2023), or asking children to freely label emotional faces or vignettes (e.g., Widen & Russell, 2008; Streubel et al., 2020). However, these approaches have limitations. For example, fluency tasks or adult-reported tasks may lack sensitivity to the development of full 'adult-like' comprehension of emotion words, as no explicit definition is required. Further, expressive free-labelling tasks may only capture a small number of emotion words, and misrepresent children's true level of emotion comprehension. Accordingly, the present chapter adopted a receptive interview-based task put forward be Nook et al. (2020). This task extends established interview-based approaches to vocabulary assessment such as the Wechsler Adult Intelligence Scale (Wechsler, 1999), with participants asked to define a set of emotion words, and responses scored based on their level of comprehension. This allows exploration of a wide range of emotion words with varying developmental trajectories - heightening sensitivity to individual differences.

4.2.3 The Current Study

While previous theoretical and empirical research points to associations between children's emotion specific language and emotion recognition accuracy for both facial and

instrumental stimuli, this association has not been examined for vocal stimuli. Accordingly, the extent to which individual differences in musical and vocal emotion recognition covary, and whether covariation can be explained by shared developmental mechanisms at the cognitive level, is unknown. Greater understanding of these mechanisms would strengthen a cross-condition model of audio emotion recognition development, which could underpin future interventions.

The current chapter explored associations between emotion recognition for instrumental, singing, and prosody stimuli in a sample of 4-8-year-old children with highly varying levels of socioemotional difficulties. This age range is characterised by increases in both emotion specific vocabulary, and general developments in emotion understanding, making it a particularly sensitive period of socioemotional development (Baron-Cohen et al., 2010; Collins, 1984; Nook et al., 2020). Individual differences in general receptive language ability and emotion language comprehension were considered as possible predictors of emotion recognition accuracy, and any condition-specific effects assessed. Given the lack of previous testing, psychometric properties for the adopted language comprehension measure were also assessed. Given possible interaction between higher-level language processing and low-level acoustic integration (Shablack & Lindquist, 2019), findings from an exploratory analysis exploring interactions between emotion language comprehension and sensitivity to acoustic features are also presented in Appendices S-T.

4.2.3.1 Hypotheses.

1. In line with evidence with TD children (Chapter 3) and adults (Chapter 4), emotion recognition accuracy for instrumental, singing, and prosody stimuli will be positively correlated. Given shared processing mechanisms for audio conditions (Escoffier et al., 2013; Frühholz et al., 2016; Paquette et al., 2018), and the high levels of individual

differences in the current sample, these correlations will be present even when accounting for age.

2. Individual differences in emotion language comprehension will predict emotion recognition accuracy, even after controlling for general verbal ability. This prediction aligns with developmental theories positing that language-mediated conceptual emotion understanding plays a central role in emotion recognition across modalities (Barrett, 2006; Cespedes-Guevara & Eerola, 2018).
3. The effect of emotion language comprehension on emotion recognition accuracy will be consistent across audio conditions. Emotion-specific language has been shown to predict musical emotion recognition in young children beyond general verbal fluency (Plate et al., 2022). This specific effect of emotion language has also been shown for facial stimuli (Streubel et al., 2020). Given the domain-generality of predictions regarding the role of emotion language in emotion development (Lindquist, 2017; Shablack & Lindquist, 2019), this association was expected to extend to vocal stimuli.

4.3 Methods

4.3.1 Participants

Children (N = 142) aged 4-8 years (M = 6.49), and their parent/guardian took part in the study. There were 82 males and 60 females. Sample sizes are noted for each individual analysis, due to variation in missingness. Participants were not deaf or hard of hearing and spoke English fluently. TD children were recruited via mainstream primary schools (n = 62, M age = 6.42, 33 female, 29 male). These TD children were also included in the sample for Chapter 3 (i.e., they are all the children included in Chapter 3 who were within the age-group adopted in the present Chapter). These children had no diagnosed developmental condition, nor scored ‘very high’ for either internalising or externalising difficulties on the Strengths and Difficulties Questionnaire (SDQ – Goodman, 1997). These very high scores are most

strongly related to a later diagnosis of a neurodevelopmental condition (Goodman et al., 2010). The average SDQ score for this group was 21.06 (SD = 5.75). Participants with a standardised score on the BPVS-3 below 70 were also excluded from this group, as this indicates a significant language difficulty (Dunn & Dunn, 2009). A further group of children were recruited via referral by their school to the Cardiff University Neurodevelopment Assessment Unit (NDAU) due to socio-emotional or behavioural difficulties (herein 'referred' sample - n = 80, M age = 6.56, 27 female, 53 male). The average SDQ score for this groups was 21.06 (SD = 5.75). Overall, 24 participants had at least half a year of formal music training, while all other participants had none.

An a priori power analysis was conducted to estimate required sample size for the linear mixed model (LMM) assessing whether emotion language comprehension predicted emotion recognition accuracy, and whether this differed between conditions. This was done using the *simr* R package (Green & MacLeod, 2016). Simulated data and effect sizes for fixed and random effects were estimated based on Chapter 3, and past research (Nook et al., 2020; Streubel et al., 2020). Models were then simulated with 100 repetitions to assess power to detect a meaningful effect at a power of 0.9, with an alpha value of .05. Findings suggested that a sample size of 100 would detect a meaningful effect for the main fixed effect of emotion language comprehension. Accordingly, the sample size of 142 (125 for this specific analysis) was deemed sufficient. The study was granted ethical approval by the Cardiff University School of Psychology Research Ethics Committee.

4.3.2 Materials and Procedure

Participants completed the tasks in a quiet room either at their school or at the NDAU testing centre, one-to-one with a researcher. The emotion recognition task was created and presented on Gorilla experiment software (Anwyl-Irvine et al., 2020). Task instructions and response options were presented in both written and verbal formats. Participants provided

their responses verbally and the researcher controlled the pace of progression through the task. Participants listened to the stimuli through JBL JR460 headphones with active noise cancellation. The language tasks were conducted using physical stimuli, while response options were recorded by the researcher on an iPad or laptop. Each testing session lasted approximately 45 minutes. Parents/guardians also completed a set of questionnaires which took approximately 10-15 minutes to complete.

4.3.2.1 Emotion Recognition Task.

4.3.2.1.1 Materials.

Twenty stimuli for each audio condition (instrumental, singing, prosody) were adopted – four for each emotion anger, fear, sadness, happiness, and calmness. Stimuli were matched between conditions as far as possible. All stimuli were high intensity and 3-7 seconds in length, in line with the 'mental presence time' for audio stimuli (Argstatter, 2016). A random set of vocal prosody and singing stimuli were taken from the validated RAVDESS stimuli set (Livingstone & Russo, 2018). Instrumental piano pieces for each emotion were taken from a range of validated stimuli sets (Micallef Grimaud & Eerola, 2021; Sutcliffe et al., 2017; Vieillard et al., 2008). For more details on stimuli selection and properties, see Chapter 2 (section 2.2.2.1.1).

Stimuli were normalised as far as possible within and between conditions (see Appendix A for details). Hardware volume was set to a normalised level across all participants. However, to ensure this sound level was not uncomfortable, an initial 'sound check' was conducted (see Chapter 3, section 3.3.2.1.1 for details). Those participants that edited the sound level were recorded, and initial analyses including/excluding these participants were run to ensure no differences in results.

4.3.2.1.2 Procedure.

Children were first introduced to the task aim – to collect coins with Shaun the Sheep by selecting the emotion word that best matched the emotion expressed by either a music clip, singing voice, or speaking voice. Trials involved the verbal and written presentation of the question 'what emotion is the music/voice/singing?', before the audio clip was triggered by the researcher. Response options then appeared on screen and were read out by the researcher, before participants verbally selected from the target emotions within a forced-choice format. For each audio condition, participants completed an initial practice phase, consisting of one trial for each emotion. Within this phase, participants received feedback on the intended emotion expression. After the practice phase, 20 experimental trials were presented for each condition. This provided an overall emotion recognition accuracy score, and one for each audio condition. For further task details, including a figure showing the visual presentation of the task, see Chapter 3 (section 3.3.2.1.2)

Audio conditions were presented separately and in a random order, and counterbalanced. Stimulus emotion within each condition were presented in an order A/order B quasi-random order (see Appendix L), counterbalanced across participants. Individual stimuli for each emotion were randomly allocated within this order, for each participant.

4.3.2.2 General Receptive Vocabulary.

4.3.2.2.1 Materials.

General receptive vocabulary was assessed via the standardised British Vocabulary Picture Scale third edition (BPVS-3; Dunn & Dunn, 2009). The BPVS-3 is a normed measure of receptive verbal ability in typically developing children and children with disabilities, with strong internal reliability (Dunn & Dunn 2009). The measure also displays strong criterion validity through correlations with the Wechsler Intelligence Scale for Children (Dunn &

Dunn, 2009; Hannant, 2018), while further evidence supports its use as a proxy for IQ in adults (Ezard et al., 2022).

4.3.2.2.2 Procedure.

Participants were presented with a set of four pictures and asked to identify the picture that matches a target word, presented verbally by the assessor. Testing followed standardised testing procedures (Dunn & Dunn, 2009). There is a maximum score of 168, organised into 14 increasingly complex sets of 12 items. Raw scores were used within analyses, but descriptive data for age-standardised scores were also presented for reference purposes.

4.3.2.3 Emotion Language Comprehension.

4.3.2.3.1 Materials.

To assess emotion language comprehension, an abridged version of Nook et al.'s (2020) emotion comprehension task was adopted. The emotion words used were the emotions from the emotion recognition task, plus two additional emotions subordinated to each of these emotion categories (subordinate as conceptualised in Shaver et al., 1987). This gave a set of 15 emotion words – angry, scared, sad, happy, calm, nervous, worried, excited, content, frustrated, embarrassed, proud, disappointed, relaxed, and jealous. These words span all four quadrants of Russel's (1980) circumplex model of emotion based on continuous planes of valence and arousal. They are also a mixture of emotion words comprehended in young children (e.g., happy, sad), and those for which comprehension emerges later (e.g., proud, frustrated - Nook et al., 2020; Baron-Cohen et al., 2010). While this task also requires expressive language skills, the emotion-specific component is receptive, and the authors liken the test to well-established measures of general receptive vocabulary such as the British Vocabulary Picture Scale (BPVS – Dunn & Dunn, 2009).

4.3.2.3.2 Procedure.

Participants were asked to define an emotion word ('what does … mean?'). Answers were audio recorded and responses scored in terms of each participants' comprehension of the emotion word (scored 0 – no comprehension; 1 – correct valence but overly vague; or 2 – full comprehension). This provided an overall emotion language comprehension score for each participant. The 15 emotion words were presented to participants in a random order. For each word, if participants did not score a maximum two points, they were probed using a set of predetermined follow-up questions. Emotion comprehension scores were assigned considering both initial and probed responses. The standardised testing protocol for this task, as described by Nook et al. (2020), was followed by all testers, as was the scoring protocol. Detailed information on this protocol can be seen in Appendix M

4.3.2.3.3 Psychometric Properties.

Psychometric assessment of the adopted abridged version of Nook et al.'s (2020) emotion language comprehension measure was conducted. An expanded sample of 179 participants aged 4-8 years (M = 6.39; 80 female; 99 male) provided data for psychometric property tests of the emotion language comprehension variable, to allow for increased confidence in psychometric assessments. This sample included all participants that completed the emotion language comprehension task, irrespective of whether they completed other tasks. Cronbach's Alpha and McDonald's Omega each indicated good internal consistency of the emotion language comprehension measure (α = 0.84; ω = 0.85). Brysbaert (2024) advises that individual items within any accuracy-based measure should have a minimum item-rest correlations coefficient of 0.2. All emotions other than 'content' (r = 0.06) met the minimum threshold of 0.2 for item-rest correlations. Further, average comprehension accuracy for this emotion (M = 0.03) was substantially lower than the average for all other emotion words (M = 1.53, *Range* = 0.89 – 1.91). Accordingly, responses to the emotion word 'content' were

removed prior to analyses. Three independent scorers scored 20% of the emotion language comprehension audio recordings. The intra-class correlation coefficients showed good absolute agreement between the three scorers, based on a two-way random effects model (kappa = 0.82, p <.001).

To ensure the measure could be appropriately conceptualised as a single score between 0 and 28, exploratory factor analysis was adopted, and scree plots visually examined. There was strong support for a single-factor solution for emotion language comprehension, with a clear drop between factor 1 and 2, and factor 2 weight in line with what would be expected from random simulated or resampled data (see Appendix N). This single factor explained 45% of the variance, and factor loadings varied from 0.40 to 0.86. Appendix O provide mean scores, standard deviations, item-rest correlations, and factor loadings for all emotion words. Test criterion and convergent validity were well supported by correlations with emotion recognition scores and other language measures, as can be seen in correlation results below.

4.3.2.4 Parent Questionnaires.

Parents/guardians provided demographic information for their child, indicated whether their child had any diagnosed developmental condition, and whether their child was deaf or hard of hearing. They also indicated their child's level of formal music training, in terms of years of instrumental or singing lessons beyond the school classroom. The validated SDQ (Goodman, 1997) was also completed to provide support that children recruited from mainstream primary schools were typically developing, and to provide an estimate of variation in socio-emotional difficulties across the sample.

4.3.3 Statistical Analysis

Frequentist statistics and BFs were included for analyses where possible. BFs provide direct evidence for both the alternative and null hypotheses, important for the present research which aims to quantify both the presence and absence of effects. BFs were interpreted in line with Jeffery's (1998) specifications, outlined in Appendix D. Inferences were based on coherence between *p*-values and BFs (e.g., evidence against an effect included a *p*-value >.05 and a BF <1/3), and any incoherence was discussed. All analyses were conducted in R (R Core Team, 2022). Full details of statistical procedures, including assumption tests, specific R packages, and reporting, can be seen in Appendix E.

4.3.3.1 Frequentist Analyses.

4.3.3.1.1 Associations Between Emotion Recognition Accuracy and Language Variables.

Associations between emotion recognition accuracy variables, language variables, and demographic variables were initially analysed via bivariate correlations. To rule out the possibility that correlations between accuracy and language variables were a function of shared associations with age, partial correlations were then conducted, while controlling for age.

4.3.3.1.2 Language Variables Predicting Emotion Recognition Accuracy.

To assess the possible contribution of general and emotion specific language to emotion recognition accuracy, and variation in these relationships between conditions, a series of linear mixed models (LMMs) were fit. Each model had a random intercept for participant. The dependent variable for each model was emotion recognition accuracy. Model 1 assessed the contribution of age, sex, condition, and BPVS scores to emotion recognition accuracy. Model 2 included emotion language comprehension as a further variable, while

model 3 also included an interaction between emotion language comprehension and condition.

4.3.3.2 Bayesian Analyses.

BFs were included for a) all correlation analyses, b) all LMMs/GLMMs at the level of fixed effects and post-hoc comparisons. They were not included for mixed model random effects, due to a lack of default priors to allow accurate inferences. For each analysis, Bayesian models were fit with fixed and random effects structures equivalent to their frequentist counterparts, using default priors (Morey et al., 2015, Oberaur, 2019, 2023). Further details of Bayesian analyses, including prior specifications, can be seen in Appendix E.

4.4 Results

4.4.1 Descriptive Statistics

Table 4.1 outlines means and standard deviations for all key variables. These are presented for the sample as a whole, and for each group (TD, referred) individually. The TD group had significantly higher emotion language comprehension scores, and higher total, instrumental, and singing emotion recognition accuracy scores, than the referred group. They also had higher standardised BPVS scores.

Table 4.1 - Means (Standard Deviations) for Whole Sample, TD Group, and Referred Group, with Group Differences

		Group			
	Maximum score	**All**	**TD**	**Referred**	**Group diff – *t* (BF^{10})**
Language variables					
BPVS Raw	168	89.6 (18.7)	90.8 (17.8)	88.8 (19.4)	0.63 (0.23)
BPVS Standardised	142	97.5 (13.2)	100.0 (11.5)	95.5 (14.1)	2.07* (1.14)
Emotion comp.	28	20.6 (5.23)	21.9 (4.48)	19.3 (5.57)	2.88** (7.39)
Emotion recognition accuracy (%)					
Instrumental	100	52.3 (50.0)	57.2 (49.5)	48.5 (0.50)	3.02** (9.79)
Singing	100	43.0 (49.5)	45.2 (49.8)	41.2 (49.2)	1.90* (0.91)
Prosody	100	56.3 (49.6)	58.6 (49.3)	54.4 (49.8)	1.59 (0.54)
Total	100	50.5 (50.0)	53.7 (49.9)	48.1 (50.0)	2.74** (4.51)

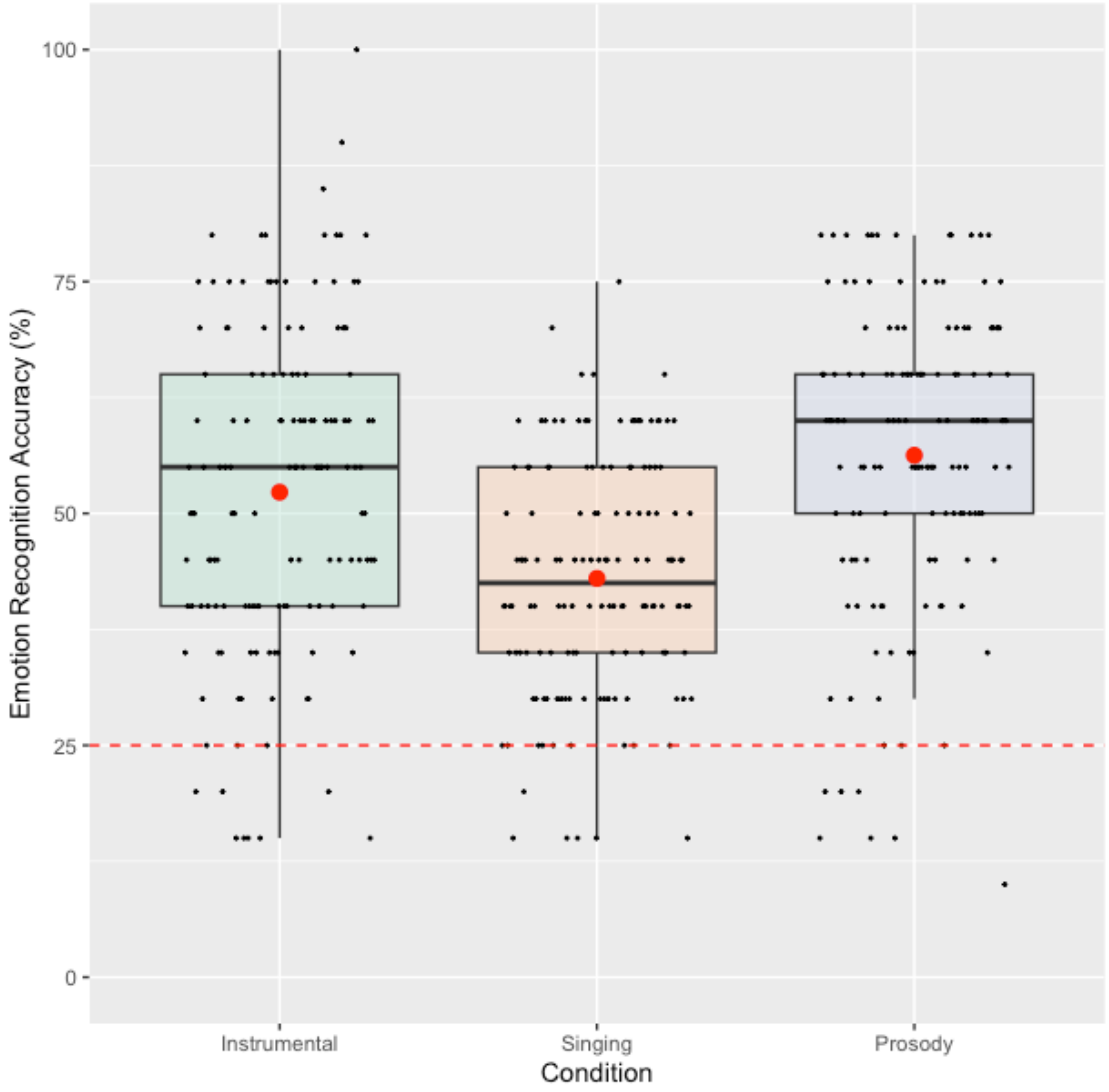
Note. N = 142 for BPVS scores and emotion recognition accuracy; 125 for emotion comprehension scores. BPVS = British Vocabulary Picture Scale; TD = typically developing. p<.05*, p<.01**, p<.001***.

4.4.2 Distribution of Emotion Recognition Accuracy Scores

Figure 4.1 displays the distribution of emotion recognition accuracy scores across the whole sample. Average scores were significantly above chance for all conditions (ps <.001; BF_{10s} >100). Both prosody and instrumental music were more accurately recognised than singing stimuli (ps <.001; BF_{10s} >100). Prosody was also more accurately recognised than

instrumental stimuli (t (141) = 2.93, p =.003, BF_{10} = 5.54). Scores were widely distributed for all conditions, but this was most pronounced for instrumental music.

Figure 4.1 – Distribution of Emotion Recognition Accuracy Scores with Means, By Condition



Note. N = 142. Red dot = condition mean. Dashed line = chance performance.

4.4.3 Associations Between Emotion Recognition Accuracy, Age, Sex, Music Training, and Language Variables

Table 4.2 outlines coefficients and BFs for correlations between emotion recognition accuracy, language variables, age, sex, and years of music training. There were relatively strong correlations between emotion recognition accuracy variables. Emotion recognition accuracy, both overall and for each condition, was also correlated with emotion language comprehension. The same was true for general receptive vocabulary, but correlations were slightly weaker (although still strongly supported by the BFs). All recognition accuracy and language variables were also correlated with age, while females had higher overall and instrumental emotion recognition accuracy than males. Females also had higher emotion comprehension scores. Finally, there was a significant correlation between emotion language comprehension scores and years of music training. There was anecdotal or moderate evidence against a correlation between music training and all other variables. Correlations conducted within each group (TD, referred) broadly aligned with these results (see Appendix P).

Table 4.2 – Whole Sample Bivariate Correlations Between Emotion Recognition Accuracy, Emotion Language Comprehension, Receptive Vocabulary (BPVS), Age, Sex, and Years of Music Training.

Variable	1	2	3	4	5	6	7	8
1. Accuracy Total	-	-	-	-	-	-	-	-
2. Accuracy Instrumental	-	-	-	-	-	-	-	-
3. Accuracy Singing	-	.46*** (>100)	-	-	-	-	-	-
4. Accuracy Prosody	-	.55*** (>100)	.48*** (>100)	-	-	-	-	-
5. Emotion Comp.	.58*** (>100)	.44*** (>100)	.44*** (>100)	.51*** (>100)	-	-	-	-
6. BPVS	.46*** (>100)	.42*** (>100)	.27*** (>100)	41*** (>100)	.65*** (>100)	-	-	-
7. Age	.35*** (>100)	.26** (22.13)	.26** (22.84)	.26*** (20.31)	.60*** (>100)	.55*** (>100)	-	-
8. Music Training (Years)	.07 (0.27)	.06 (0.26)	.04 (0.22)	.05 (0.23)	.28* (16.03)	.06 (0.25)	.06 (0.26)	-
9. Sex	-.19* (2.43)	-.18* (1.62)	-.19* (2.10)	-.11 (0.47)	-.22* (6.07)	-.10 (0.37)	-.09 (0.39)	-.05 (0.23)

Note. Coefficient (BF_{10}). Spearman's rho correlations involving music training and age. Biserial correlations involving sex. Pearson's correlations between all other variables. *df* = 123 between variables and emotion language comprehension; 142 between all other variables. All variable scores are raw (unadjusted). BPVS = British Vocabulary Picture Scale. $p<.05$*, $p<.01$**, $p<.001$***.

Table 4.3 outlines partial correlations between emotion recognition and language variables, while controlling for age. Emotion recognition accuracy variables remained relatively strongly correlated. Total accuracy, and accuracy for each condition, also remained significantly correlated with language variables, but as above, these were generally stronger for emotion language comprehension compared to general receptive vocabulary. For singing, although significant, the partial correlation between singing recognition accuracy and BPVS scores was not supported by the BF, which provided anecdotal evidence against this relationship.

When the sample was analysed as two separate groups, correlation direction and strength aligned with correlations within the referred-only sample. Results also broadly aligned with the TD sample, but correlations were generally weaker, and BPVS scores were only related to emotion recognition accuracy for instrumental music, for this sub-sample (see Appendix Q for full correlation tables).

Table 4.3 – Whole Sample Partial Correlations Between Emotion Recognition Accuracy, Emotion Language Comprehension, and Receptive Vocabulary (BPVS), Controlling for Age

Variable	1	2	3	4	5
1. Accuracy Total	-	-	-	-	-
2. Accuracy Instrumental	-	-	-	-	-
3. Accuracy Singing	-	.45*** (>100)	-	-	-
4. Accuracy Prosody	-	.51*** (>100)	.42*** (>100)	-	-
5. Emotion Comp.	.48*** (>100)	.34*** (>100)	.34*** (>100)	.43*** (>100)	-
6. BPVS	.36*** (>100)	.34*** (>100)	.16* (0.56)	.34*** (>100)	.47*** (>100)

Note. Coefficient (BF_{10}). Pearson's correlations. *df* = 123 between variables and emotion language comprehension; 142 between all other variables. All variable scores are raw (unadjusted). BPVS = British Vocabulary Picture Scale. $p<.05$*, $p<.01$**, $p<.001$***.

4.4.4 Do Individual Differences in General and Emotion-Specific Language Predict Emotion Recognition Accuracy?

A series of LMMs were fit to test the effect of language variables on emotion recognition accuracy. Model 1 assessed the contribution of general verbal ability (BPVS) on emotion recognition accuracy while accounting for age, sex, and condition. Model 2 added emotion language comprehension as a further variable to test the hypothesised unique effect of emotion-specific language, while model 3 also included an interaction between emotion language comprehension and condition, to test hypothesised consistency in this effect between conditions.

As Table 4.4 shows, while controlling for age, sex, and condition, model 1 provided strong evidence to suggest that BPVS scores significantly positively predicted emotion

recognition accuracy. There were also significant, but anecdotal, effects of sex and age, and a significant effect for condition. The model's fixed effects explained 25% of the variance in emotion recognition accuracy.

In model 2, there was moderate evidence *against* an effect for BPVS scores, while there was a significant positive effect of emotion language comprehension on emotion recognition accuracy, strongly supported by the BF. There was also anecdotal evidence against an effect for sex, and moderate evidence against one for age, in this model. Model 2 better predicted emotion recognition accuracy, explaining 31% of variance.

Model 3 provided evidence against an interaction between emotion language comprehension and condition, moderately supported by the BF. The main effect for emotion language comprehension, and evidence against an interaction between emotion language comprehension and condition, was mirrored within the referred and TD samples, independently, as can be seen in Appendix R.[14]

[14] The one divergence from these results was a lack of main effect for BPVS scores for the TD sample, in model 1.

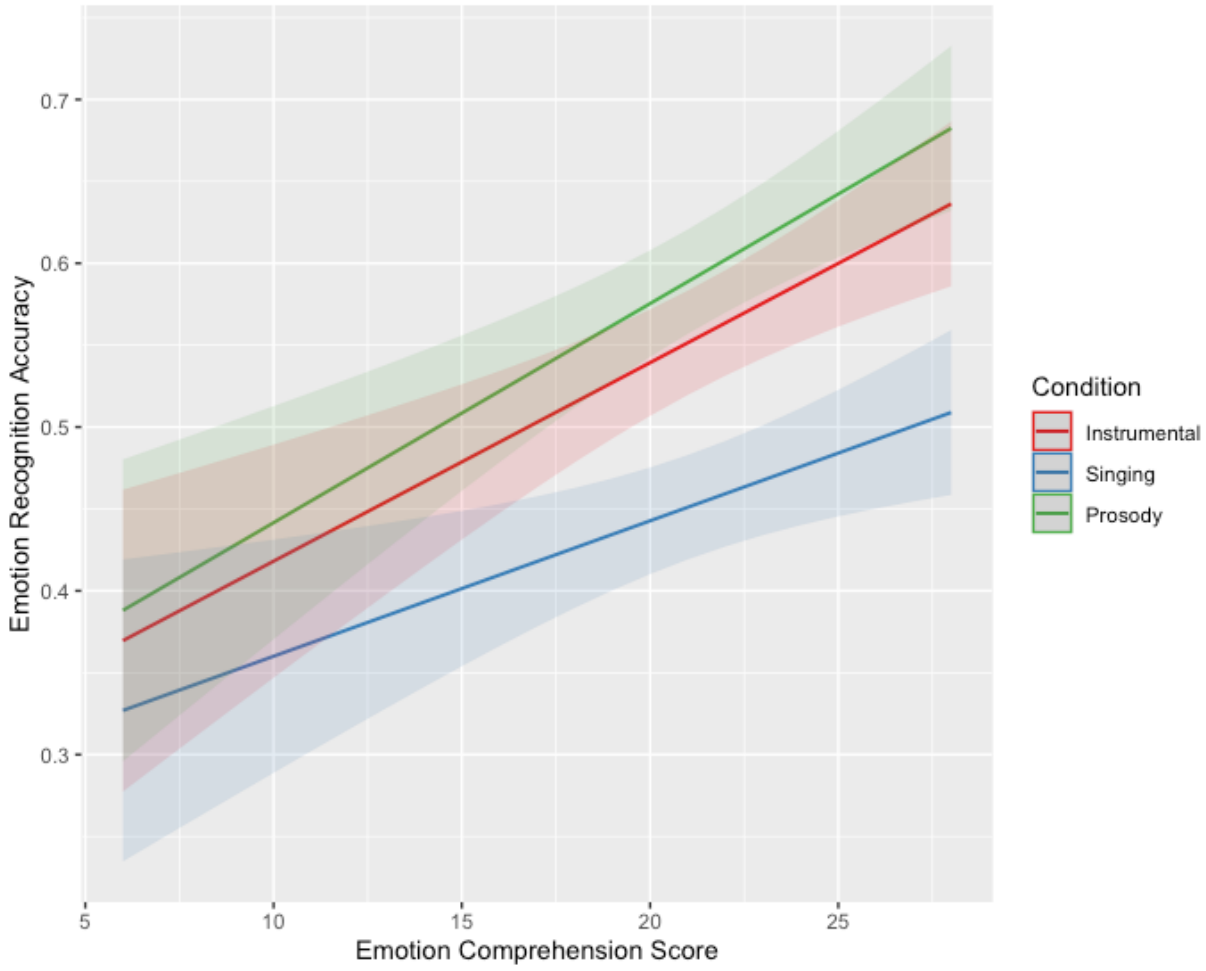
Table 4.4 - Fixed Effects and Marginal R Squared for Linear Mixed Models on Emotion Recognition Accuracy

	DV: Emotion Recognition Accuracy	
	X^2	BF_{10}
Model 1		
Intercept	6.44**	
Age	4.17*	1.33
Sex	4.48*	1.64
Condition	94.61***	>100
BPVS	8.48**	11.00
Model Marginal R^2	.25	
Model 2		
Intercept	9.65**	
Age	0.29	0.21
Sex	1.86	0.47
Condition	94.61***	>100
BPVS	0.34	0.22
Emotion Comp.	21.52***	>100
Model Marginal R^2	.31	
Model 3		
Intercept	9.65***	
Age	0.29	0.28
Sex	1.86	0.49
Condition	95.28***	>100
BPVS	0.40	0.28
Emotion Comp	21.52***	>100
Emotion Comp*Condition	3.76	0.14
Model Marginal R^2	.32	

Note. $p<.05$*, $p<.01$**, $p<.001$***. $N = 125$. BPVS = British Vocabulary Picture Scale.

Based on model 3 results, Figure 4.2 presents the marginal effect of emotion language comprehension scores on emotion recognition accuracy by condition, while controlling for age, sex, and BPVS scores. It shows a clear difference in overall recognition accuracy between conditions, with prosody (M = 0.57, SE = 0.01) and instrumental music (M = 0.53, SE = 0.01) each recognised more accurately than singing stimuli (M = 0.44, SE = 0.01, ps <.001, BF_{10s} >100). However, the strength of effect of emotion language comprehension on recognition accuracy is relatively similar between conditions.

Figure 4.2 - Marginal Effect of Emotion Language Comprehension on Emotion Recognition Accuracy, by Condition



Findings from the exploratory analysis of interactions between participants' emotion language comprehension and acoustic feature levels, in relation to the likelihood that each emotion would be selected, are presented in Appendices S-T.

4.5 Discussion

The current chapter extended past research by exploring associations between emotion recognition for instrumental, singing, and prosody stimuli in a sample of 4-8-year-old children with varying levels of socioemotional difficulties. Emotion language comprehension was considered as a possible condition-general predictor of individual differences in emotion recognition accuracy. Findings strengthen theoretical and developmental links between musical and vocal emotion understanding, with strong positive correlations between emotion recognition accuracy for all audio conditions. Findings also indicated that emotion language comprehension, over and above general language abilities, predicted individual differences in emotion recognition accuracy across audio conditions. This aligns with research stressing the importance of language-mediated emotion concept development for emotion recognition (Shablack & Lindquist, 2019), and points to a possible condition-general mechanism of audio emotion recognition development.

Positive correlations between emotion recognition accuracy for instrumental, singing, and prosody stimuli were strong with the present sample of children with highly varied levels of socioemotional difficulties. This remained the case when accounting for age and held within the referred sample (higher levels of socioemotional difficulties) in isolation. This extends past research demonstrating shared developmental trajectories for instrumental and vocal emotion recognition (Heaton & Allgood, 2015; Vidas et al., 2020), and positive associations between accuracy for these conditions and singing (Chapter 3), in typical

development. Findings add robustness to the links between instrumental and prosodic emotion recognition, in that even within this more varied sample, recognition ability for music and voice appears to covary. When considered alongside research highlighting some shared expressive cues (Juslin & Laukka, 2003; Llie & Thompson, 2006), and neurophysiological processing mechanisms (Escoffier et al., 2013; Proverbio et al., 2020), findings further substantiate a theorised functional basis to the development of musical and vocal emotion recognition (Juslin, 2018; Scherer, 1995; 2017; Brown, 2017). However, despite the present correlations between accuracy for audio conditions, there was no evidence for a relationship between music training and emotion recognition accuracy for any audio condition, aligning with past chapters. This supports the possibility that musical aptitude may be a key acoustic-perceptual mechanism, operating independent of formal music training (Vigl et al., 2024: Jansen et al., 2023), that partially explains associations between emotion recognition accuracy for musical and vocal stimuli.

The strong positive association between emotion recognition accuracy for singing and other audio conditions, but greater relative difficulty recognising sung emotions overall, aligns with Chapters 2 and 3 of the present thesis. This may support the proposal that the ability to draw meaning from singing forms an early basis for the later developing ability to comprehend musical and vocal emotions (Boone & Cunningham, 1998; Flom & Bahrick, 2007; Trehub, 2001), and for socio-emotional development more broadly (Schubert & McPherson, 2015; Politimou et al., 2018). Indeed, in young children, singing is more effective than speech at drawing attention to the mouth (Alviar et al., 2023), and ambiguous stimuli is more likely to be categorised as song rather than speech (Vanden Bosch der Nederlanden et al., 2022). Further, Chapter 3 indicated that emotion recognition accuracy for singing stimuli predicted accuracy for both instrumental and prosody conditions in TD children. This was the case while controlling for age, music training, and emotion recognition

accuracy for prosody and instrumental music, respectively. This suggests a fundamental developmental role for the ability to draw emotional meaning from singing stimuli - supported by the present study's strong positive correlations between singing and both instrumental and prosodic emotion recognition accuracy in a sample of children with varied levels of socioemotional difficulties.

Shared expressive cues and processing mechanisms at the acoustic-perceptual level offer only a partial explanation for strong correlations between accuracy for audio conditions, as audio emotion recognition also involves a higher-order process of cognitive interpretation (Bestelmeyer et al., 2014). Accordingly, the present chapter hypothesised that language ability may be a possible developmental mechanism of individual differences in musical and vocal emotion recognition accuracy. Findings supported this hypothesis. There were positive associations between general receptive language ability and emotion recognition accuracy, across conditions, while controlling for age and sex. This aligns with past research finding associations between children's general language ability and facial emotion recognition accuracy (Beck et al., 2012; Pons et al., 2003; Rosenqvist et al., 2014; Streubel et al., 2020), as well as a range of other aspects of emotion development (for review, see Shablack & Lindquist, 2021). Importantly, they are also congruent with longitudinal evidence for a directional association between language ability and vocal emotion recognition development (Griffiths et al., 2020), as well as the vocal emotion recognition difficulties experienced by children with DLD (Boucher, 2000; Griffiths et al., 2020; Taylor et al., 2015). The present findings provide further, direct evidence for associations between general language abilities and vocal emotion recognition accuracy, within a sample of children with varied levels of socioemotional difficulties. They also support analogous links between general language and emotion recognition in musical stimuli, in line with past research (Franco et al., 2017). This is unsurprising, given music's inextricable relationship with the development of language

abilities in early life (Franco et al., 2022; Pino et al., 2023; Politimou et al., 2019), and extensive links between language and emotion processing (Shablack & Lindquist, 2021).

Importantly, in the present chapter, when considered alongside emotion language comprehension, there was evidence *against* an association between general verbal ability and emotion recognition accuracy. Meanwhile, emotion language comprehension strongly predicted individual differences in emotion recognition accuracy, suggesting a specific and unique association between emotion language comprehension and audio emotion recognition accuracy. This held across audio conditions, while controlling for age and sex, and remained the case within the referred sample in isolation. The measure of general language adopted in the present research – the BPVS (Dunn & Dunn, 2009) – has been strongly correlated with IQ in children (Hannant, 2018), and within some samples, shown to be a suitable proxy for its measurement (Ezard et al., 2022). Given links between IQ and vocal emotion recognition ability (Jones et al., 2011; Schlegel et al., 2020), and the correlations between BPVS scores and emotion recognition accuracy for all audio conditions within the present study, findings evidence a unique, cross-condition association between emotion language comprehension and emotion recognition accuracy.

The specific role of emotion language aligns with past research stressing the importance of emotion language comprehension for emotion recognition development - extending them to the vocal domain. Strong links between domain-specific vocabulary and the development of conceptual understanding in that domain have been demonstrated (e.g., Theory of Mind – Grazzani & Ornaghi, 2012). For emotion development, conceptual understanding of emotions may enable children to organise affective experiences and perceptual information within specific emotion categories (Barrett, 2006, 2017; Lindquist, 2017) and develop more mature emotional skills (Nenchevca et al., 2023; Nook et al., 2020).

Importantly, some of these emotion theories centralise language as fundamental to the ability to make meaning from more basic perceptual properties of stimuli (Barrett, 2006, 2017; Lindquist, 2017). Indeed, the priming or impeding of emotion concepts via words affects the way in which adults perceive and experience emotions (Nook et al., 2015; Satpute et al., 2016), including the way they perceive low-level perceptual features (Brooks & Freeman, 2018). This would indicate a more interactive relationship between the acoustic integration and cognitive interpretation stages of Bestelmeyer et al.'s (2014) model. While an exploratory analysis in the present chapter (Appendix T) indicated that higher levels of emotion language comprehension may influence how certain acoustic features relate to emotion perceptions, the correlational analytic approach does not allow for firm inferences. Future research could elucidate interactions between acoustic-perceptual and language-based cognitive processing stages during musical and vocal emotion perceptions, via word priming/impeding paradigms.

Regardless of the processing stages at which it influences emotion recognition, the development of emotion language comprehension appears important to audio emotion recognition. This is reflected in the developmental literature. Associations have been demonstrated between children's emotion-specific language and facial emotion recognition accuracy, understanding of mixed emotions, and knowledge of emotional experiences/expressions in 7-9-year-olds (Beck et al., 2012). Further, certain associations, including with facial emotion recognition accuracy, also operate over and above the effect of general verbal ability for 4-9-year-old children (Streubel et al., 2020). The present findings align with and extend these findings, implicating emotion language comprehension as a possible mechanism for the development of vocal emotion recognition. However, Grosse and Streubel (2024) found that while expressive emotion-specific vocabulary, rather than general expressive vocabulary, predicted adolescents' facial emotion recognition accuracy, the opposite was true for 10-11-year-old children. Accordingly, future research should consider a

wider age range of children to explore the possibility of a complex interplay between the development of general language, emotion-specific language, and vocal emotion recognition accuracy.

The current findings also support the presence of unique associations between emotion language comprehension and emotion recognition for more abstract (or at least less directly adaptive) forms of emotion expression, singing and instrumental music. This suggests a possible condition-general link between emotion-specific language and audio emotion recognition in 4-8-year-old children. The consistency of this association for musical and vocal stimuli aligns with research indicating similar processing mechanisms for musical and other forms of emotion stimuli (Lima et al., 2016; Escoffier et al., 2013; Proverbio et al., 2020a; Proverbio et al., 2020b; Proverbio et al., 2022). In this vein, Fritz et al. (2019) suggest that music can 'stimulate representations of semantic content that are shared with the verbal modality', as indicated by its influence on novel word learning. Indeed, music and voice appear to activate similar ventral regions of the primary auditory cortex involved in the conceptual station of stimuli (Schirmer et al., 2012). As noted, emotion language is propounded as a scaffold for perceptual emotion development (Barrett, 2006, 2017; Satpute & Lindquist, 2021; Shablack & Lindquist, 2019) via facilitating conceptual understanding of emotions, including their causes, consequences, and forms of expression (Lindquist, 2017; Nook et al., 2017). Plausibly, then, music's similarities to the voice in terms of the expression and processing of meaning may facilitate more accurate emotion recognition for individuals with more advanced emotion language comprehension. Indeed, Plate et al. (2022) found that expressive emotion language fluency predicted individual differences in musical emotion recognition accuracy, while a measure of general verbal fluency did not. Overall, the present findings suggest that alongside commonalities in expressive and perceptual processing mechanisms for musical and vocal stimuli (Escoffier et al., 2013; Jentschke, 2016; Juslin &

Laukka, 2003; Proverbio et al., 2020), emotion-specific language ability may be one condition-general mechanism of audio emotion recognition development. This association could partially explain similarities in the development of emotion recognition between conditions, despite some condition-specific accuracy levels for certain emotions (Heaton & Allgood, 2015; Vidas et al., 2020; Chapter 3).

The present Chapter also provides support for the use of an abridged 14-item version of Nook et al.'s (2020) emotion language comprehension task. While it is likely that the full version would provide the best approximation of emotion language comprehension, developmental research is often limited by the need to balance analyses involving numerous constructs with the maintenance of adequate statistical power (Bergmann et al., 2018; Davis-Kean & Ellis, 2019). Inter-rater reliability, internal consistency, and test validity of the shortened version were good, while exploratory factor analysis supported a single-score conceptualisation of emotion language comprehension ability. The task adopted has limitations, however. While it offers a purely linguistic measure of comprehension, performance could be limited by either participants' receptive understanding of the word, *or* their ability to express its meaning (Streubel et al., 2020; Sturrock & Freed, 2023). Future research could control for general expressive ability or adopt tasks that isolate these receptive and expressive components (e.g., Sturrock & Freed, 2023; Streubel et al., 2020). However, aspects of these tasks carry their own limitations, such as some reliance on visual emotion recognition, and less flexibility to explore a wider range of emotion words. Given a growing research base highlighting the role of emotion-specific language within emotion development (Beck et al., 2012; Lindquist et al., 2015; Ornaghi & Grazzani, 2013; Streubel et al., 2020), it is positive that a wider range of measures are emerging. However, future research should be clear regarding the study-specific justifications for their measure of choice, and, particularly

in the case of research with non-typical samples, conscious of the sample-specific nature of any previous reliability assessments (Brysbaert, 2024).

4.5.1 Limitations and Future Directions

The present study also had some further limitations. First, the cross-sectional nature of the sample limits inferences regarding directional association between emotion language and emotion recognition development. Indeed, for music stimuli, given proposed causative associations between musical environment and later language development (Papadimitriou et al., 2021; Politimou et al., 2019), the association between emotion-specific language and musical emotion recognition may be multi-directional. Future longitudinal research could elucidate the directionality of developmental associations between emotion language and emotion recognition (see Griffiths et al., 2020). A second limitation relates to the low overall level of music training in the sample. Accordingly, inferences relating to possible acoustic-perceptual mechanisms such as musical aptitude, beyond formal music training (Vigl et al., 2024), would be strengthened with a sample of children with higher overall levels of music training. As discussed, the present methods were not well-suited to exploring the mechanism through which emotion language comprehension may influence audio emotion recognition. Future research could also explore interactions between acoustic/low-level perceptual factors – generally proposed as central to similarities in how music and voice are expressed, processed, and perceived (Juslin & Laukka, 2003; Scherer, 2017) – and conceptual/language-based aspects of emotion recognition development. This would bring developmental research more in line with much ongoing research for visual stimuli (e.g., Woodward et al., 2019), and with theoretical advances stressing a fundamental role for emotion concepts within emotion perception development (Shablack & Lindquist, 2019).

4.5.2 Conclusion

Overall, the present findings extend past research highlighting condition-general audio emotion recognition mechanisms at the acoustic-perceptual level (e.g., Vigl et al., 2024) - highlighting one possible mechanism at the cognitive level in emotion language comprehension. Associations between the development of this ability and audio emotion recognition appears sensitive across children with a range of socio-emotional difficulties, and consistent between audio conditions. This aligns with research underlining general socio-emotional processing mechanisms for musical and vocal emotions beyond the acoustic-perceptual level (Lima et al., 2016; Escoffier et al., 2013) and supports theoretical conceptualisations that propound shared functional roots for the expression and perception of musical and vocal emotions (e.g., Juslin, 2018; Brown, 2017). For developmental research, the present findings indicate a need to further explore condition-general and condition-specific mechanisms underpinning audio emotion recognition development. This would allow for a deeper understanding of the ways in which vocal and musical emotion recognition converge and diverge across development, and the nature/implications of individual differences in these processes. Importantly, given success of musical interventions in improving children's vocal emotion recognition and a range of broader socio-emotional abilities (Blasco-Magraner et al., 2021), incorporation of language-based components highlights an avenue through which the efficacy of interventions could be enhanced.

5. Associations Between Vocal and Musical Emotion Recognition and Socio-Emotional Adjustment in Children

5.1 Relationship with Previous Chapters

Chapters 2 and 3 found broad similarities in adult's and TD children's emotion perception patterns for prosody, singing, and instrumental stimuli. However, there were some condition-specific patterns, as well as some differences in the associations between stimulus-level features and emotion perceptions. Chapter 4 built on these findings, exploring individual differences in emotion recognition patterns in children with a wide range of socioemotional difficulties. Language development was considered as a possible condition-general mechanism of emotion recognition development. Findings indicated that emotion-specific language, rather than general verbal ability, predicted individual differences in emotion recognition accuracy, across audio conditions.

The present chapter shifted focus to exploring associations between vocal and musical emotion recognition accuracy, and socioemotional dimensions, in children referred by their school due to socioemotional, cognitive or behavioural difficulties exhibited at school. Understanding which dimensions are related to emotion recognition accuracy, and whether they are similar or differ between musical and vocal stimuli, could be informative for future research and for music-based interventions targeting socioemotional dimensions.

5.2 Introduction

5.2.1 Vocal Emotion Recognition, and Internalising and Externalising Behaviours/Difficulties

Child socioemotional development is often conceptualised within two broad dimensions – externalising and internalising behaviours. Externalising behaviour refers to

behaviours directed outwards to the social environment, including those relating to attention/hyperactivity, and aggressive, oppositional, and rule-breaking behaviours. Internalising behaviours have an internal focus, such as anxious and depressive symptoms, withdrawal, and somatic complaints (Achenbach & Edelbrock, 1978). Higher scores for children on each of these dimensions is related to poorer outcomes later in development (Bor et al., 2004; Korhonen et al., 2018; Papchristou & Flouri, 2020). This focus on externalising and internalising difficulties aligns with a transdiagnostic approach, which emphasises underlying dimensions of functioning rather than categorical diagnoses (Eaton et al., 2015). This perspective may provide a more ecologically valid understanding of how emotion recognition ability relates to behaviour across populations with diverse developmental profiles (Insel et al., 2010) – as is relevant for the present chapter's referred sample.

Despite extensive links between vocal emotion processing and social competence in children (Nowicki & Mitchell, 1998; Scheerer et al., 2020; Verbeek, 1996), research examining links between vocal emotion recognition and internalising/externalising difficulties is relatively limited. However, evidence points more consistently to an association with externalising problems, relative to internalising problems. Indeed, Chronaki et al. (2015a) found that vocal prosody emotion recognition accuracy was negatively correlated with externalising, but not internalising, difficulties in a group of 3-6-year-old children with and without behavioural problems. Similarly, Nowicki et al. (2019) found that vocal emotion recognition accuracy at age 8 was associated with total behavioural and emotional difficulties at age 10 in a sample of TD children. However, this study did not parse externalising and internalising problems to explore any independent relationships. Conversely, some research with TD children has found no relationship between vocal emotion recognition and either internalising or externalising difficulties, although relationships were apparent for overall social competence and other dimensions including behavioural and cognitive self-regulation

(Neves et al., 2021; Nowicki & Mitchell, 1998). More extensive research involving facial expressions indicates that associations between emotion recognition and the externalising dimension are stronger in groups of children with overall higher levels of externalising difficulties (Cooper et al., 2020). This may partially explain the null findings above regarding an association between vocal emotion recognition and externalising difficulties in TD children. Seemingly, while some evidence points to an association between externalising difficulties and vocal emotion recognition accuracy, inconsistencies remain. While these can be partially attributed to sample differences, consideration of specific dimensions within this broader externalising category can clarify this picture.

5.2.2 Considering Specific Externalising Dimensions

Externalising is often broken down into behavioural and attentional aspects. Research exploring associations between behavioural aspects and vocal emotion recognition development is limited but indicates a possible association. For example, a meta-analysis indicated an emotion-general negative association between psychopathic traits and vocal emotion recognition accuracy, in both children and adults (Dawel et al., 2012). While an extreme case, psychopathic traits are strongly related to externalising problems and aggressive behaviour in TD and clinical samples of children (Demetriou et al., 2023; Grossi et al., 2023; Muratori et al., 2021). However, conversely, in a sample of 8-12-year-old children with high levels of rule-breaking behaviour, there was no evidence for a difficulty with vocal emotion recognition relative to TD controls, despite difficulties with facial emotion recognition for this group (van Zonnevald et al., 2019).

In the case of attentional aspects, there is also some evidence for associations with vocal emotion recognition accuracy. For example, a recent meta-analysis found evidence for vocal emotion recognition difficulties in children with ADHD (Sells et al., 2023). Further, research with a clinical sample indicated that children with ADHD or CD each had

difficulties with emotion recognition (in terms of a combined facial and vocal accuracy score) relative to TD children (Cadesky et al., 2000). Interestingly, when considered together, Chronaki et al. (2015a) found that hyperactivity, over and above conduct problems, predicted vocal emotion recognition accuracy in 3-6-year-old children. Although the specific mechanisms of these associations are unclear, this may suggest specific links between attentional processes/hyperactivity and vocal emotion recognition accuracy. This is unsurprising, given the need to attend to and integrate complex acoustic information during vocal emotion recognition judgements (Bestelmeyer et al., 2014). However, more recent facial expression research indicated that co-occurring CD explained any ADHD-related difficulties with facial emotion recognition, while eye-tracking data suggested that perceptual rather than attentional mechanisms underpinned these difficulties (Airdrie et al., 2018). Emotion recognition of faces and voices activates some distinct areas of the brain (Schirmer & Adolphs, 2017) and some vocal research has indicated atypical attentional processing of emotions in children with ADHD even when those with co-occurring CD are excluded (Chronaki et al., 2018). However, it is possible that more extensive vocal expression research accounting for co-occurring conditions would highlight a more prominent role for conduct-related difficulties interpreting emotion stimuli as key to links between externalising problems and vocal emotion recognition accuracy (Airdrie et al., 2018).

Seemingly, although there is some evidence to suggest associations between children's vocal emotion recognition accuracy and both behavioural and attentional aspects of the externalising dimension, evidence is mixed. Much past research has not involved children with higher levels of socioemotional difficulties. Those that did have either focused on more general emotion recognition ability rather than vocal emotions (e.g., Cadesky et al., 2000), explored traits such as psychopathy rather than externalising and related sub-dimensions (e.g., Dawel et al., 2012), or included only a relatively small sub-sample of children with

behavioural problems (Chronaki et al., 2015a). Accordingly, research that maximises individual differences in socioemotional dimensions while considering both broader internalising and externalising dimensions as well as sub-dimensions therein may add clarity to the research base. Alongside these limitations, there is also a lack of research extending beyond facial and vocal domains to include other forms of emotion expression, such as music. Music and voice are closely linked in terms of the expression of emotion (Eerola et al., 2013; Grandjean et al., 2006; Juslin & Laukka, 2003), neural and behavioural processing mechanisms (Escoffier et al., 2013; Proverbio et al., 2020; Vigl et al., 2024; Chapter 4), and the development of emotion recognition (Vidas et al., 2018; Heaton & Allgood, 2015; Chapter 3). Given further links between broader musical abilities and socioemotional development, a consideration of musical and vocal emotion recognition in tandem in relation to socioemotional development could be a fruitful avenue for research.

5.2.3 Music and Socioemotional Development

A growing research base links a range of musical abilities and various aspects of socioemotional development. In a recent systematic review, Blasco-Magraner et al. (2021) found a robust association between the educational use of music and children's emotion perception and regulation. These links were apparent for formal music training, and in relation to specific musical interventions. For example, children with a musical background exhibit higher prosocial skills and sympathy (Schellenberg et al., 2015). Further, Rose et al. (2015) found that children spending more time learning a musical instrument had lower externalising problems, conduct problems, hyperactivity, aggression, and attention problems, compared to those spending less time learning. In terms of musical interventions, Boucher et al. (2021) found that 4-5-year old's participation in a music programme focused on a range of musical skills (through experimenting with music and movement in a variety of styles, tonalites, etc) produced increases in social skills and independence, while older children also

improved in their emotion comprehension. Further, children receiving additional musical instrument training displayed lower physical and verbal aggression, as well as an improved ability to perceive emotions in the face, relative to children receiving only classroom music lessons (Kim & Kim, 2018). Interestingly, no such effects were apparent for anxiety, or for overall emotion understanding. Similarly, there is some evidence linking music therapy with positive socioemotional outcomes. Gold et al.'s (2004) meta-analysis indicated that music therapy has a positive and substantial impact on socioemotional outcomes for children and adolescents - particularly for behavioural outcomes, relative to emotional ones.

Despite these general benefits of music for socioemotional development, few studies have examined *emotional* understanding of music, specifically, in relation to socioemotional development. One exception found that TD adolescents' ability to perceive musical emotions and express them via manipulation of a range of acoustic cues related to their self-reported levels of empathy and conduct problems. Specifically, findings indicated associations between the incongruent expression of emotions sadness and anger through music, and levels of externalising problems. There were also associations between cognitive and affective empathy and the perception and expression of certain emotions (Saarikallio et al., 2014). This may indicate that adolescents' musical emotion understanding mirrors to some extent their broader abilities in socioemotional communication. It is possible that further associations with perceptual aspects of musical emotion understanding would emerge with different sample parameters, such as with younger children with more varied socioemotional difficulties. Overall, although associations between musical abilities/training and various socioemotional dimensions, particularly those related to the externalising dimension, are apparent, there is a lack of research exploring *emotional* understanding of music in relation to these dimensions. When considered on these terms, possible commonalities between music and the voice emerge.

5.2.4 Vocal and Musical Emotion Recognition – Converging and Diverging Mechanistic Associations with Socioemotional Development

Past research has indicated correlations between emotion recognition accuracy for music and voice for adults and children both with and without socio-emotional difficulties (Chapter 2; Chapter 3; Heaton & Allgood, 2015; Vidas et al., 2018). These associations may be underpinned by condition-general emotion processing mechanisms at the acoustic-perceptual level and cognitive level (Escoffier et al., 2013; Proverbio et al., 2020; Vigl et al., 2024; Chapter 4). Accordingly, one avenue through which musical abilities may relate to socio-emotional dimensions could be via an improved comprehension of emotional aspects of music, and its associations on these terms with the voice. In line with this possibility, music training is positively associated with vocal emotion recognition abilities (Martins et al., 2020; Nussbaum & Schweinberger, 2021). Further, an intervention focused specifically on learning how emotions in music are expressed lead to improvements in adults' ability to recognise emotions in the voice, while an art-based intervention with an equivalent aim/structure did not (Mualem & Lavidor, 2015). Interestingly, this study also found no difference between participants with long-term formal music training and those without in their ability to recognise vocal emotions - pin-pointing emotional understanding of music as key to improvements in vocal emotion recognition. Given these commonalities, as well as the links between more general musical abilities and socioemotional dimensions discussed, it could be expected that musical and vocal emotion recognition would relate similarly to socioemotional dimensions. Specifically, it could be hypothesised that there would be common associations with externalising, but less so internalising, dimensions, in children with socioemotional difficulties (Chronaki et al., 2015a).

However, past research also illuminates potential condition-specific relationships between emotion recognition and socioemotional dimensions. For example, although

attention mechanisms may be key to vocal emotion recognition (Sells et al., 2023), it is unclear whether such mechanisms would be relevant for musical stimuli. Indeed, while there is evidence to suggest that music is processed by similar mechanisms as employed for social stimuli, including the voice (Bedoya et al., 2021; Escoffier et al., 2013), it is also the case that music is more strongly self-referential – it directs attention to the acoustic material itself, while language often refers to something else in the environment (Reybrouck & Podlipniak, 2019). As prosody is generally accompanied by language, this may indicate a difference in how the attentional processes involved in musical emotion processing develop, in relation to more explicitly social stimuli such as the voice. However, the implications this may have in relation to these attentional mechanisms, and their role in the links between emotion recognition and externalising problems, is unclear. Indeed, as discussed, the role of attention mechanisms in relation to externalising-related emotion recognition difficulties for facial stimuli appears limited (Airdrie et al., 2018), while suggestions regarding the role of attention mechanisms for vocal stimuli are based on a relatively small research base (Sells et al., 2023).

There are also some music-specific processing mechanisms that may suggest independent associations with socio-emotional development. For example, music-specific associations with socio-emotional development, and externalising problems specifically, could be linked to children's early musical environment (see Politimou et al., 2018). Early in development, this primarily involves infant-directed song. Infant-directed song effectively modulates arousal (Shenfield et al., 2003), is rated as more emotional than speech (Trehub et al., 2016) and heightens interest and attention towards the caregiver (Theissen & Saffran, 2009). This musical environment may have implications for relationships between musical emotion recognition and socioemotional dimensions. For example, musical engagement between caregiver and child supports positive attachment and emotional connection (Creighton et al., 2013; Fancourt & Perkins, 2018). Caregiver-child attachment, particularly

higher levels of insecure or avoidant attachment, is positively associated with externalising and internalising problems in children and adolescents (Achenbach et al., 2016). Further, children's early musical environment has been shown to relate to children's emotion regulation later in development (Williams et al., 2015). Given the integral role of emotion dysregulation in externalising problems (Mullin & Hinshaw, 2007), it could be that children's musical environment has implications for the development of externalising difficulties. Importantly, though, these assertions rest on the untested assumption that home musical environment relates to higher musical emotion recognition accuracy. Further, the specificity of these links is contingent on any associations between musical emotion recognition accuracy and socioemotional dimensions being independent of those with vocal emotion recognition accuracy. Given the proposal that musical and vocal emotion recognition develop from a *shared* basis in the music-like interactions of early development (Schubert & McPherson, 2015), it may be more likely that musical and vocal emotion recognition display similar associations with socioemotional dimensions.

5.2.5 The Current Study

Despite relationships between various musical abilities and socioemotional dimensions, no research has examined associations between musical emotion recognition and socioemotional development in children. Further, research considering these links for vocal emotions is limited, particularly with samples of children with higher levels of socioemotional difficulties. Such a sample could be more sensitive to shared variance in both socioemotional dimensions and emotion recognition accuracy, while providing findings applicable to children who are most likely to benefit from intervention. Considering musical and vocal emotion recognition in tandem would be informative regarding the form such interventions could take – which socioemotional dimensions could be targeted and whether they should focus on shared or condition-specific developmental mechanisms. Given the

apparent importance of singing to the links between early musical environment and child development, consideration of both instrumental and singing stimuli could also provide unique insights.

Accordingly, the current chapter explored associations between emotion recognition accuracy for vocal and musical (instrumental, singing) stimuli, and socioemotional dimensions, in 5-8-year-old children referred by teachers due to socioemotional, cognitive or behavioural difficulties. Externalising and internalising dimensions were analysed initially before an explorative analysis involving some externalising-related sub-dimensions.

5.2.5.1 Hypotheses.

1. Vocal prosody emotion recognition will be negatively correlated with externalising, but not internalising, difficulties. This prediction was based on evidence for more consistent relationships between externalising, relative to internalising, difficulties and both vocal and facial emotion recognition (Chronaki et al., 2015a; Cooper et al., 2020).
2. Given overlapping neural and perceptual mechanisms for musical and vocal emotion processing (Escoffier et al., 2013; Juslin & Laukka, 2003; Vigl et al., 2024), strong correlations in recognition accuracy between these conditions (Chapter 3; Chapter 4; Vidas et al., 2018; Heaton & Allgood, 2015), and independent relationships between various musical abilities and externalising-related difficulties (Kim & Kim, 2018; Rose et al., 2015), instrumental and singing emotion recognition will display the same pattern as vocal prosody.

Given the limited amount of research, further analyses with specific dimensions will be exploratory in nature.

5.3 Methods

5.3.1 Participants

Eighty-four children aged 5-8 years (M = 6.56) and their parent/guardian took part in the study. There were 55 males and 29 females. Sample sizes are provided for each analysis due to differences in missingness. Participants were not deaf or hard of hearing and spoke English fluently. Participants were referred by their school to the NDAU at Cardiff University, due to socioemotional, behavioural, or cognitive difficulties exhibited at school. These were the same children who were also included as the 'referred' sample for Chapter 4. 14 participants had at least half a year of formal music training, while all others had none. Gross household income was categorised on a 7-point Likert scale. The median household income value reported was 5 (£40,000 - £49,000), ranging between 1 (up to £9,999) and 7 (£60,000+).

An a priori power analysis conducted with the *pwr* R package (Champley et al., 2022) indicated that 83 participants would be required to achieve 80% power to detect a moderate correlation ($r = 0.3$), with an alpha value of 0.05. Due to missingness, some analyses did not reach this threshold - the limitations of this are discussed. The study was granted ethical approval by the Cardiff University School of Psychology Research Ethics Committee.

5.3.2 Materials and Procedure

Child participants completed the emotion recognition task one-to-one with a researcher in the NDAU testing centre. The task was designed and presented on Gorilla experiment software (Anwyl-Irvine et al., 2020). Task instructions and response options were presented in both written and verbal formats. Participants provided their responses verbally and the researcher controlled the pace of progression through the task. Participants listened to

the stimuli through JBL JR460 headphones with active noise cancellation. The task lasted approximately 10-15 minutes.

Parents/guardians provided sociodemographic information for their child before the session, while they completed a questionnaire on their child's behavioural and emotional problems with a researcher at the NDAU testing centre. In total, these components took approximately 15-20 minutes.

5.3.2.1 Emotion Recognition Task.

5.3.2.1.1 Materials.

Sixty audio emotion stimuli – 20 for instrumental, singing, and prosody conditions – were adopted. They expressed emotions anger, fear, sadness, happiness, and calmness. All stimuli were between 4-7 seconds in length, in line with the 'mental presence time' for audio stimuli (Augstatter et al., 2016). A random set of vocal prosody and singing stimuli were taken from the validated RAVDESS stimuli set (Livingstone & Russo, 2018). Instrumental music stimuli were piano pieces taken from a range of databases (Micallef Grimaud & Eerola, 2021; Sutcliffe et al., 2017; Vieillard et al., 2008). For more details on stimuli selection and properties, see Chapter 2 (section 2.2.2.1.1).

Stimuli were normalised as far as possible within and between conditions (see Appendix A for details). Hardware volume was also normalised across participants. However, an initial 'sound check' was conducted, to ensure the sound level was not uncomfortable (see Chapter 3, section 3.3.2.1.1 for details).

5.3.2.1.2 Procedure.

Children were first introduced to the overall aim of the task – to select the emotion word that best matched the instrumental, singing, or vocal prosody clip. For each audio condition, participants completed an initial practice phase, consisting of one trial for each

emotion. Within this phase, participants received feedback on the intended emotion expression. For each trial (20 per audio condition), the sentence ‘what emotion is the music/singing/voice?’ was presented visually and verbally to the participant, before the researcher triggered the audio clip. Response options were then presented visually and read aloud by the researcher, and participants provided their responses verbally within a forced-choice format. An overall emotion recognition accuracy score, and one for each audio condition, was produced for each participant. For further task details, including a figure showing the visual presentation of the task, see Chapter 3 (section 3.3.2.1.2).

Audio conditions were presented separately and in a random order, and counterbalanced. Stimulus emotion within each condition were presented in an order A/order B quasi-random order (see Appendix L), counterbalanced across participants. Individual stimuli for each emotion were randomly allocated within this order, for each participant.

5.3.2.2 Parent Questionnaires.

Before the session, parents/carers provided sociodemographic information, including their child’s age, sex, and years of formal music training (instrumental or singing). They also indicated their level of gross household income on a 7-point Likert scale (1 = up to £9,999; 2 = £10,000 - £19,999; 3 = £20,000 - £29,999; 4 = £30,000 - £39,999; 5 = £40,000 - £49,999; 6 = £50,000 - £59,999; 7 = £60,000+).

During the session, parents/carers completed the Child Behaviour Checklist (CBCL – Achenbach & Rescoria, 2001) to assess their child’s socioemotional and behavioural difficulties. The CBCL is a 113-item measure, with each item representing a statement about the child scored on a 3-point Likert scale from 0 = ‘not true’, to 2 = ‘certainly true’. Children are then scored along a range of socioemotional and behavioural dimensions. These dimensions reflect syndrome scales, DSM-V oriented scales, and broad composite scores

denoting internalising and externalising difficulties. The present study focused on internalising and externalising scores, as well as a further set of externalising-related dimensions (Aggressive Behaviour, Rule-Breaking Behaviour, Oppositional-Defiant Problems, Attention Problems). Raw scores were used within statistical analyses, but descriptive data also included standardised t-scores, and the proportion of participants falling within 'borderline' or 'clinically significant' categories. Scores within these categories suggest the need for more comprehensive diagnostic assessment (Pandolfi et al., 2012).

5.3.3 Statistical Analysis

Both Frequentist and Bayesian analyses – in the form of Bayes Factors (BFs) – were included. BFs were included as they provide an index of strength of evidence for both the null and the alternative hypothesis. BFs were interpreted in line with Jeffery's (1998) specifications, outlined in Appendix D. Inferences were based on coherence between *p*-values and BFs (e.g., evidence against an effect included a *p*-value >.05 and a BF <1/3), and any incoherence was discussed. All analyses were conducted in R (R Core Team, 2022). Full details of statistical procedures, including assumption tests, specific R packages, and reporting, can be seen in Appendix E.

For frequentist analyses, associations between emotion recognition accuracy and CBCL dimensions, as well as demographic variables, were initially analysed via bivariate correlations. Due to a lack of normality within CBCL variables, Spearman's rho correlations were adopted. To rule out the possibility that relationships between emotion recognition and socioemotional dimensions were a function of shared associations with age, partial correlations were also computed, controlling for age. Some significant associations were then analysed via multiple regression, to assess the contribution of emotion recognition accuracy to socioemotional dimensions, while controlling for age, sex, music training, household income, and accuracy for other relevant audio conditions. Models were fit with robust

standard errors, due to the presence of outliers. Bayesian analyses mirrored the above, adopting the default priors within the BayesFactor R package (Morey et al., 2015). Further details on Bayesian analyses, including prior specifications, can be seen in Appendix E.

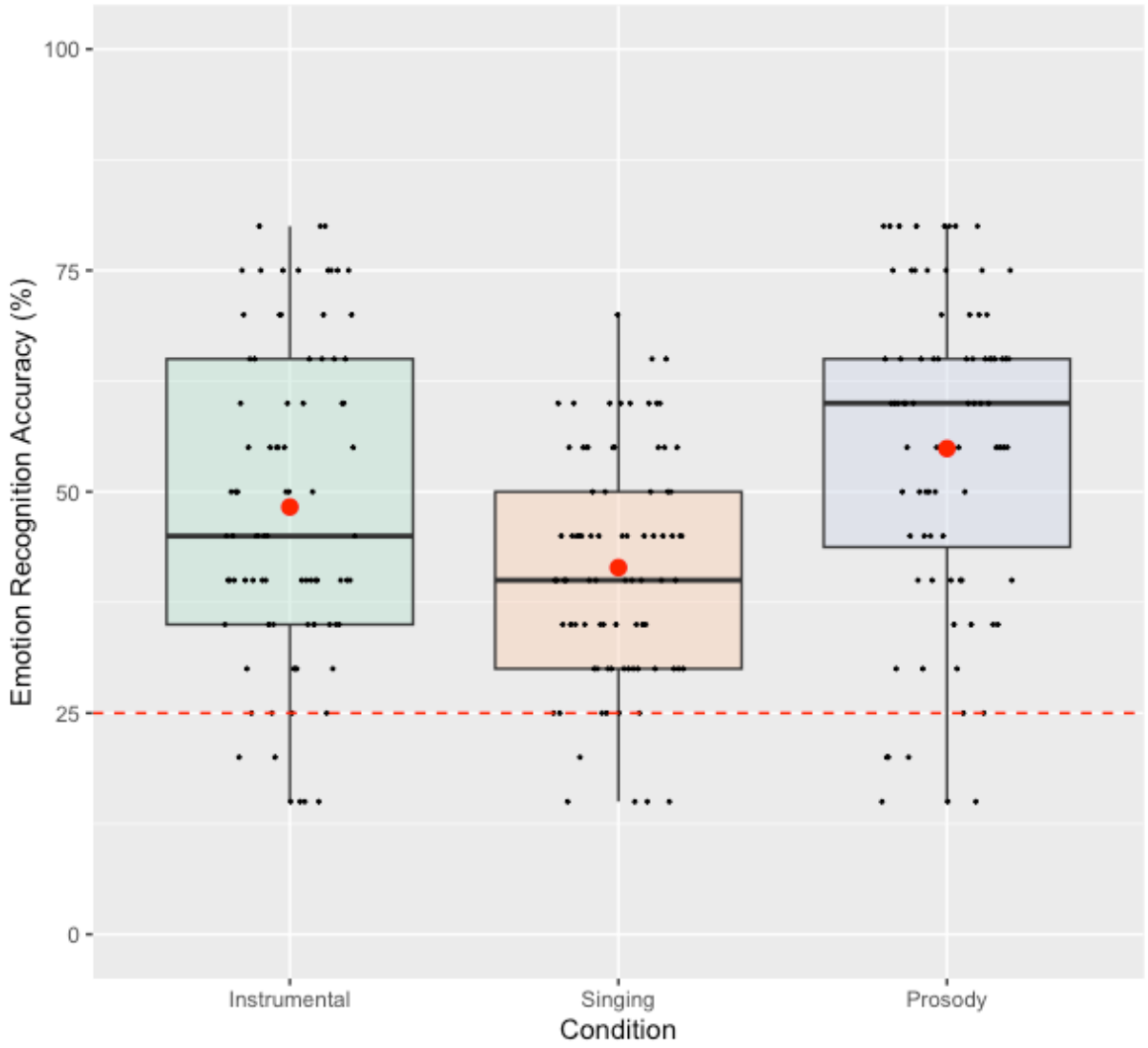
5.4 Results

5.4.1 Descriptive Statistics

5.4.1.1 Emotion Recognition.

Figure 5.1 outlines distributions of emotion recognition scores for instrumental, singing, and prosody conditions, as well as mean scores and chance accuracy level. Average scores for all conditions were significantly above chance (*p*s <.001; BF_{10s} >100). Both instrumental music and prosody stimuli were more accurately recognised than singing stimuli, and prosody was more accurately recognised than instrumental stimuli, overall (*p*s <.001; BF_{10s} >100). Distributions were relatively wide for all conditions, but this was most pronounced for instrumental music stimuli.

Figure 5.1 – Distribution of Emotion Recognition Accuracy Scores with Means, by Condition



Note. N = 84.

5.4.1.2 Socioemotional Dimensions and Socio-demographic Variables.

Table 5.1 outlines means, SDs, and ranges for raw socioemotional dimension variable scores. It also outlines t-scores for these variables, and the proportion of participants whose t-scores fell into borderline and clinically significant categories.

Table 5.1 - Means and Standard Deviations for Socioemotional Dimensions (CBCL)

	CBCL Dimension					
	Externalising	**Internalising**	**Aggressive**	**Rule Breaking**	**Oppositional**	**Attention**
Raw mean (SD)	21.18 (10.54)	14.05 (10.16)	15.89 (8.36)	4.81 (2.92)	5.92 (2.80)	9.56 (4.31)
Range	3 - 45	0 - 41	3 - 35	0 - 11	0 - 12	1 - 19
T-score (SD)	65.77 (9.93)	60.75 (11.60)	67.0 (11.49)	62.50 (7.58)	63.99 (8.94)	68.93 (10.35)
% Clinically sig.	62	43	38	25	40	45
% Borderline	10	11	19	19	11	25

Note. $N = 52$ for rule breaking; 73 for all other dimensions. CBCL = Child Behaviour Checklist. Clinically sig. = clinically significant.

5.4.2 Associations Between Emotion Recognition Accuracy and Socioemotional Dimensions

Table 5.2 outlines bivariate correlations between emotion recognition accuracy, externalising and internalising dimensions, music training, and sociodemographic variables. There were relatively strong positive correlations between emotion recognition accuracy variables. Overall accuracy and instrumental accuracy were also negatively correlated with externalising, but there was moderate evidence against one for internalising. Externalising and internalising were also each negatively correlated with age and household income, while total recognition accuracy and prosody recognition accuracy were each positively correlated with age.

Table 5.2 – Bivariate Correlations Between Emotion Recognition Accuracy, Internalising and Externalising Difficulties, and Sociodemographic Variables

Variable	1	2	3	4	5	6	7	8	9
1. Accuracy Total	-	-	-	-	-	-	-	-	-
2. Accuracy Instrumental	-	-	-	-	-	-	-	-	-
3. Accuracy Singing	-	.46*** (>100)	-	-	-	-	-	-	-
4. Accuracy Prosody	-	.56*** (>100)	.56*** (>100)	-	-	-	-	-	-
5. Externalising	-.29** (5.19)	-.36*** (26.52)	-.21^ (1.25)	-.14 (0.49)	-	-	-	-	-
6. Internalising	-.03 (0.29)	-.02 (0.27)	.09 (0.37)	-.06 (0.30)	.44*** (>100)	-	-	-	-
7. Music Training	.06 (0.29)	.01 (0.25)	.13 (0.50)	.04 (0.27)	-19 (0.85)	-.16 (0.59)	-	-	-
8. Age	.22* (1.56)	.07 (0.31)	.20^ (1.22)	.27* (4.59)	-.26* (3.99)	-.34** (14.17)	.10 (0.37)	-	-
9. Sex	.11 (0.40)	.01 (0.25)	.17 (0.74)	.12 (0.42)	.09 (0.36)	.29* (16.17)	.02 (0.25)	.01 (0.25)	-
10. Family Income	.12 (0.44)	.11 (0.41)	.10 (0.37)	.05 (0.28)	-.35* (14.57)	-.43** (>100)	.28* (4.36)	.03 (0.26)	-.15 (0.57)

Note. Coefficient (BF_{10}). Pearson's correlations between accuracy variables and age. Spearman's rho correlations involving music training, and internalising/externalising due to lack of normality. Point-biserial correlations involving sex. $df = 82$ for correlations between accuracy variables, age, and sex; 71 for correlations involving externalising/internalising. All scores = raw scores. p <.10^, p<.05*, p<.01**, p<.001***

Table 5.3 outlines partial correlations between emotion recognition accuracy and externalising/internalising variables, while controlling for age. Recognition accuracy variables remained strongly positively correlated, while there was anecdotal evidence for a negative correlation between externalising and total accuracy, and moderate evidence for one with instrumental accuracy.

Table 5.3 – Partial Correlations Between Emotion Recognition Accuracy, and Internalising and Externalising Difficulties, Controlling for Age

Variable	1	2	3	4	5
1. Accuracy Total	-	-	-	-	-
2. Accuracy Instrumental	-	-	-	-	-
3. Accuracy Singing	-	.45*** (>100)	-	-	-
4. Accuracy Prosody	-	.57*** (>100)	.53*** (>100)	-	-
5. Externalising	-.25* (1.74)	-.35** (9.83)	-.17 (0.48)	-.07 (0.29)	-
6. Internalising	.06 (0.28)	.02 (0.27)	.17 (0.58)	.04 (0.27)	.41** (58.92)

Note. Coefficient (BF_{10}). Pearson's correlations between accuracy variables. Spearman's rho correlations involving internalising/externalising due to lack of normality. *df* = 82 for correlations between accuracy variables; 71 for correlations involving externalising/internalising. p <.10^, p<.05*, p<.01**, p<.001***.

Table 5.4 outlines bivariate correlations between emotion recognition accuracy variables, CBCL dimensions Aggressive Behaviour, Rule-Breaking Behaviour, Oppositional-Defiant Problems, and Attention Problems, as well as music training and sociodemographic variables. There was moderate evidence for a negative correlation between total recognition

accuracy and Rule Breaking Behaviour. For instrumental music, there was moderate evidence for a negative correlation with both Aggressive and Rule-Breaking Behaviour, and anecdotal evidence for one with Oppositional-Defiant Behaviour. There was anecdotal-moderate evidence against a correlation between other accuracy variables and socioemotional dimensions. Some CBCL dimensions were also correlated with age and income.

Table 5.4 – Bivariate Correlations Between Emotion Recognition Accuracy, Externalising-Related Dimensions, and Sociodemographic Variables

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Accuracy Total	-	-	-	-	-	-	-	-	-	-	-
2. Accuracy Instrumental	-	-	-	-	-	-	-	-	-	-	-
3. Accuracy Singing	-	.46*** (>100)	-	-	-	-	-	-	-	-	-
4. Accuracy Prosody	-	.56*** (>100)	.56*** (>100)	-	-	-	-	-	-	-	-
5. Aggressive	-.22^ (1.25)	-.29* (4.38)	-.17 (0.67)	-.06 (0.31)	-	-	-	-	-	-	-
6. Rule Breaking	-.35** (5.98)	-.37** (9.95)	-.22^ (0.96)	-.19 (0.68)	.79*** (>100)	-	-	-	-	-	-
7. Oppositional	-.17 (0.63)	-.24* (1.85)	-.09 (0.36)	-.05 (0.29)	.84*** (>100)	.63*** (>100)	-	-	-	-	-
8. Attention	-.01 (0.27)	.06 (0.30)	-.14 (0.51)	-.01 (0.26)	.12 (0.48)	.30* (2.79)	.02 (0.27)	-	-	-	-
9. Music Training	.06 (0.29)	.01 (0.25)	.13 (0.50)	.04 (0.27)	-.20^ (0.85)	-.17 (0.64)	-.05 (0.29)	.03 (0.28)	-	-	-
10. Age	.22* (1.56)	.07 (0.31)	.20^ (1.22)	.27* (4.59)	-.28* (2.99)	.06 (0.34)	-.25* (2.03)	.42*** (>100)	.10 (0.37)	-	-
11. Sex	.11 (0.40)	.01 (0.25)	.17 (0.74)	.12 (0.42)	.18 (0.58)	-.17 (0.61)	.16 (0.63)	-.06 (0.30)	.02 (0.25)	.01 (0.25)	-
12. Household Income	.12 (0.44)	.11 (0.41)	.10 (0.37)	.05 (0.28)	-.34** (12.08)	-.15 (0.53)	-.33** (10.92)	-.25* (1.97)	.28* (4.36)	.03 (0.26)	-.15 (0.57)

Note. Coefficient (BF_{10}). Pearson's correlations between accuracy variables and age. Point-biserial correlations involving sex. Spearman's rho for all other correlations. *df* = 82 for correlations between accuracy variables, age, sex, and income; 71 for correlations between these variables and aggression, oppositional, attention; 50 for correlations between these variables and rule breaking. p <.10^, p<.05*, p<.01**, p<.001***.

Table 5.5 outlines partial correlations between recognition accuracy and Aggressive Behaviour, Rule-Breaking Behaviour, Oppositional-Defiant Problems, and Attention Problems, while controlling for age. There remained a significant negative correlation between total recognition accuracy and Rule-Breaking Behaviour. There was moderate evidence for a negative correlation between instrumental accuracy and Aggressive Behaviour, and strong evidence for one with Rule-Breaking Behaviour. There was also moderate evidence for a negative correlation between singing recognition accuracy and attention problems.

Table 5.5 – Partial Correlations Between Emotion Recognition Accuracy and Externalising-Related Socio-Emotional Dimensions, Controlling for Age

Variable	1	2	3	4	5	6	7
1. Accuracy Total	-	-	-	-	-	-	-
2. Accuracy Instrumental	-	-	-	-	-	-	-
3. Accuracy Singing	-	.45*** (>100)	-	-	-	-	-
4. Accuracy Prosody	-	.57*** (>100)	.53*** (>100)	-	-	-	-
5. Aggressive	-.16 (0.55)	-.27* (3.31)	-.12 (0.35)	.03 (0.29)	-	-	-
6. Rule Breaking	-.35* (6.89)	-.38* (14.46)	-.22^ (0.90)	-.20 (1.07)	.79*** (>100)	-	-
7. Oppositional	-.12 (0.48)	-.22^ (1.39)	-.05 (0.28)	.03 (0.27)	.83*** (>100)	.63*** (>100)	-
8. Attention	-.13 (0.37)	.02 (0.28)	-.24* (3.15)	-.14 (0.55)	.27* (6.14)	.30* (1.32)	.14 (0.58)

Note. Coefficient (BF_{10}). Pearson's correlations between accuracy variables. Spearman's rho for all other correlations. *df* = 82 for correlations between accuracy variables; 71 for correlations between these variables and aggression, oppositional, attention; 50 for correlations between these variables and rule breaking. p <.10^, p<.05*, p<.01**, p<.001***.

5.4.3 Does Instrumental Recognition Accuracy Predict Externalising Difficulties Independent of Sociodemographic Variables?

Table 5.6 displays results from a robust multiple regression model, predicting externalising difficulties from age, sex, household income, music training, prosody recognition accuracy, and instrumental recognition accuracy. While holding all other variables constant, there was moderate evidence to suggest that instrumental accuracy negatively predicted variance in externalising difficulties. There was also moderate evidence for a significant negative effect of household income on externalising difficulties.[15]

Table 5.6 – Robust Multiple Regression Model for Externalising Difficulties

	DV: Externalising		
	β	**SE**	**BF_{10}**
Model 1			
Intercept	45.80***	8.78	
Age	-2.05	1.61	1.02
Sex	2.64	2.51	0.39
Household Income	-1.35*	0.57	3.31
Music Training	-1.66	3.10	0.43
Prosody Accuracy	6.56	8.15	0.51
Instrumental Accuracy	-19.79**	6.78	3.87
Model Adjusted R^2	.19		

Note. β = unstandardised regression coefficient. $p<.05$*, $p<.01$**, $p<.001$***. $N = 66$.

[15] A model without prosody accuracy is presented in Appendix U. The effect of instrumental accuracy was significant but slightly weaker in this model.

5.5 Discussion

The current chapter explored vocal and musical emotion recognition in relation to externalising and internalising socioemotional dimensions, in a group of children with elevated socioemotional and behavioural difficulties. Findings for instrumental music aligned with hypotheses, observing a negative association between accuracy and externalising difficulties. This appeared related to behavioural, rather than attentional, externalising sub-dimensions. For singing, the opposite pattern was true, with an association for attention-related aspects of the externalising dimension, although evidence was weaker compared to instrumental stimuli. However, contrary to expectations, there was no evidence for an association between vocal emotion recognition accuracy and externalising or any externalising-related sub-dimension. Findings support the possibility that emotional understanding of music is an important musical ability that relates to children's broader socioemotional development – specifically behavioural aspects of the externalising dimension. This appears to operate independently of any shared association with or through vocal emotion recognition, although methodological limitations that may have contributed to this finding are discussed below.

In the present chapter, prosody recognition accuracy was not significantly associated with either internalising or externalising dimensions. In the case of internalising difficulties, this aligns with past research with children of a similar age (Chronaki et al., 2015a; Neves et al., 2021; Nowicki & Mitchell, 1998). However, some studies with facial expressions found relationships between emotion recognition for certain emotions, such as sadness and fear, and internalising problems (Dede et al., 2021), and a more general difficulty has been found in children with anxiety disorders (Easter et al., 2005). Further, some research suggests that adults with anxiety disorders also have difficulties with vocal emotion recognition (Tseng et al., 2017). Accordingly, although the present sample had elevated levels of internalising

problems, further vocal prosody research with clinical samples of children could add clarity to the research base.

For externalising, the lack of relationship with vocal prosody recognition accuracy conflicts with some past research. For example, Chronaki et al. (2015a) found that 3-6-year-old's prosody recognition accuracy was significantly negatively associated with both conduct and hyperactivity aspects of the externalising dimension, aligning with research involving facial stimuli (Cooper et al., 2020). This discrepancy could be attributed to methodological differences. For example, Chronaki et al. (2015a) presented vocal stimuli at two levels of intensity. Although in the present chapter overall accuracy and score distributions did not indicate ceiling effects, it could be that the task adopted lacked sufficient sensitivity to detect relationships with the externalising dimension. Indeed, although non-significant, there was a negative trend in the relationship between prosody recognition and certain dimensions, such as rule-breaking behaviour and to a lesser extent, attention difficulties. Future cross-condition research could adopt simpler musical stimuli such as the Musical Emotion Bursts (Paquette et al., 2013) and attempt to validate low and high intensity versions of these stimuli via manipulation of salient acoustic features (e.g., loudness). This would allow direct cross-condition comparisons of relationships with socioemotional dimensions at varying levels of stimulus intensity.

The lack of association between vocal emotion recognition accuracy and externalising behaviour in the present chapter could also be linked to sample characteristics. While Chronaki et al.'s (2015a) study was with 3-6-year-old children, the present study included children aged 5-8 years. Past research with children of a similar age to the present chapter found no relationship between vocal emotion recognition accuracy and externalising problems (Neves et al., 2021). Accordingly, it may be that vocal emotion recognition accuracy most strongly relates to externalising problems earlier in development. In line with

this assertion, past research indicated most rapid development in vocal emotion recognition accuracy between 4-5 and 6-7 years (Chapter 3; Chronaki et al., 2015b). Further, research points to a shift from an auditory dominance towards a visual dominance within children's emotion perceptions during this developmental period (Ross et al., 2021; Ross et al., 2023). Taken together, this evidence suggests that the ability to recognise vocal emotions may be more strongly related to externalising difficulties early in development. Indeed, one study found that emotion recognition accuracy for facial expressions was related to externalising problems at 6-years, but not at 8-years, while there was a lack of longitudinal relationship between the constructs (Castro et al., 2018). Similar research exploring age-related differences in the links between prosody recognition accuracy and externalising difficulties, involving children both with and without developmental difficulties, could clarify this picture.

Partially in line with hypotheses, there was a significant association between instrumental recognition accuracy and externalising difficulties in the present chapter. This remained when controlling for sociodemographic variables, music training, and prosody recognition accuracy, suggesting an independent association between instrumental accuracy and externalising difficulties. While there was a similar negative trend for singing stimuli, this did not reach significance. This finding does not appear to support the possibility that musical emotion recognition accuracy relates to externalising problems via its associations with the vocal modality, in line with links between music training and vocal emotion recognition accuracy (Martins et al., 2020; Mualem & Lavidor, 2015; Nussbaum & Schweinberger, 2021). However, it may be that recognition of emotion in instrumental music – at least within the present task parameters – represents a more sensitive measure of children's ability to perceive and interpret acoustic variations in stimuli (i.e., musical aptitude). Multiple studies have implicated this ability as fundamental to both musical and

vocal emotion recognition (Vigl et al., 2024; Jansen et al., 2023). Further, some theories have conceptualised music as a 'super-expressive voice', due to its ability to exaggerate acoustic features to powerfully express and evoke emotion (e.g., Juslin, 2001). Within this context, emotional understanding of instrumental stimuli may reflect a more advanced comprehension of acoustic variations that are key to audio emotion perceptions. Indeed, instrumental music and singing stimuli were each recognised less accurately than prosody overall, while the distribution of instrumental scores was particularly wide. This heightened level of individual differences may have rendered it sensitive to variation in the externalising dimension.

However, if it were the case that this shared acoustic-perceptual mechanism was the primary explanation for the observed association between instrumental music and externalising difficulties, the inclusion of prosody recognition accuracy should have at least lessened the strength of the effect of instrumental accuracy on externalising difficulties. However, its inclusion marginally *increased* the strength of the effect of instrumental music, suggesting a suppression effect of prosody recognition accuracy (Appendix U). Accordingly, the present association between instrumental emotion recognition and externalising difficulties may reflect more general aspects of musical emotion understanding. Indeed, past research has indicated that processing of musical emotions is underpinned to some degree by general socio-emotional processing mechanisms, beyond low-level acoustic processing systems shared with the voice (Chapter 4; Escoffier et al., 2013; Lima et al., 2016). Further, research has pointed to associations between musical abilities and music therapies/interventions and various aspects of children's socioemotional development (Blasco-Magraner et al., 2021; Gold et al., 2004), including externalising problems (Kim & Kim, 2018; Rose et al., 2015). Based on the present findings, it could be that children's emotional understanding of music, as a reflection of their emotion development more generally, is one avenue through which musical abilities relate to externalising problems.

Indeed, that this relationship was apparent even when controlling for formal music training isolates understanding of music at the emotional level as key to links to externalising problems. However, this inference relies on replication in a sample of children with overall higher levels of music training, given the low overall levels and lack of independent relationship with emotion recognition seen in the present chapter.

However, the above evidence does not highlight anything *specific to music* (other than representing a more acoustically expressive form of audio stimuli) that could link it, independent of vocal prosody, to externalising difficulties in children. There are two lines of research that may elaborate these music-specific pathways. First, children's early musical experiences and environment could influence both their instrumental emotion recognition accuracy and their externalising problems. For example, evidence suggests that music engagement between caregiver and child supports positive attachment and emotional connection (Creighton et al., 2013; Fancourt & Perkins, 2018; Sanfillipo et al., 2021). Caregiver-child attachment, particularly higher levels of insecure or avoidant attachment, is positively associated with externalising and internalising problems in children and adolescents (Achenbach et al., 2016). Further, children's early musical environment has been shown to relate to children's emotion regulation later in development (Williams et al., 2015). Emotion dysregulation is integral to externalising problems (Mullin & Hinshaw, 2007), and for some groups of children, emotion recognition ability may partially mediate the link between dysregulation and externalising difficulties (e.g., those with ADHD – Graziano & Garcia, 2016; Rosen et al., 2019). Accordingly, it could be that children's early musical environment has implications for the development of the observed relationship between musical emotion recognition and externalising difficulties. However, these assertions rest on the untested assumption that a more extensive or diverse musical environment relates to higher musical emotion recognition accuracy. Further, there is a strong genetic component to

musical ability (e.g., Kragness et al., 2021; Mosing et al., 2017; Tan et al., 2014). Thus, the extent to which trait-like musical abilities and early musical environments independently and interactively relate to emotional understanding of music, and broader socioemotional dimensions like externalising difficulties, could be an important avenue for future research. Such research could illuminate the mechanisms underpinning the observed association between instrumental emotion recognition and externalising difficulties, informing targeted musical interventions.

A second music-specific mechanism that may link it independently to externalising difficulties relates to certain music-specific processing mechanisms. For example, music-specific processing mechanisms have been strongly linked to neural processing of reward (Blood & Zatorre, 2001; Zatorre & Salimpoor, 2013), while individual differences in reward sensitivity influences adults' emotion perceptions such that positive emotions are perceived as more pleasant, and some negative emotions as less pleasant, in those with high reward-sensitivity (Feuntes-Sánchez et al., 2023). Atypical reward processing has also been extensively linked to higher externalising problems in children (Gatzke-Kopp et al., 2009; Kasparek et al., 2020). Plausibly, then, differences in reward processing could relate to both instrumental recognition accuracy and externalising difficulties, partially explaining music-specific associations with externalising difficulties within Chapter 5.

The present study observed a difference in the specific externalising dimensions associated with instrumental and singing recognition accuracy. For instrumental music, associations were with aggressive and rule-breaking behaviours, while for singing, there was a negative correlation with attention problems, once age was controlled for. This discrepancy could be linked to the specific role of singing in development. As noted, infant-directed singing is a key component of very early musical environments and development (Politimou et al., 2018). Importantly, infant-directed singing heightens interest and attention towards the

caregiver (Shön et al., 2008; Theissen & Saffran, 2009) and rhythmically entrains them to a rich source of interpersonal information, supporting socio-emotional development (Lense et al., 2022). Accordingly, it may be that this early developmental role of singing leads to a later relationship between attention problems and singing emotion recognition accuracy. Instrumental musical emotion recognition, on the other hand, may reflect a more refined form of musical emotion understanding, rendering it a more likely mediator of some of the links between musical abilities/interventions and behavioural externalising outcomes discussed above (e.g., Kim & Kim, 2018).

5.5.1 Limitations and Future Directions

The present study had a range of limitations, further to those discussed above. Most analyses were slightly underpowered based on the a priori power analysis, particularly those involving the rule-breaking behaviour dimension. However, the negative correlation between instrumental recognition accuracy and rule-breaking behaviours (r = -.38) had an adequate observed power of .81, while the BF also provided strong evidence for this correlation. Still, future research with a larger sample could strengthen inferences. Further, due to a lack of previous research, the present study focused on overall recognition accuracy for each audio condition, rather than accuracy for specific emotions. Although evidence is mixed, some past research suggests associations between externalising difficulties and emotion-specific emotion recognition difficulties for facial (Airdrie et al., 2018) and vocal (Sells et al., 2023) stimuli. Future research could explore this possibility for vocal and musical stimuli in tandem, to facilitate better specified interventions. Future longitudinal research could also better elucidate the nature and directionality of the observed association between musical emotion recognition and externalising difficulties. Such research could consider early musical environments, general musical abilities, and emotional understanding of music, in relation to the development of externalising difficulties. This would facilitate understanding of the

extent to which understanding of emotional aspects of music, and/or more general musical abilities, relate to the development of externalising difficulties.

It should also be noted that the present chapter focused exclusively on *perceived* rather than *felt* emotion in response to musical and vocal stimuli. However, given the present proposal for music-specific relationships between musical emotion understanding and externalising difficulties, future research could focus more explicitly on the felt aspect of musical emotion experience. Indeed, the link between early musical environment and emotion regulation development (Williams et al., 2015), and strong associations between emotion dysregulation and externalising difficulties (Mullin & Hinshaw, 2007), may make this felt aspect of musical emotion experience a productive target for intervention with children with externalising difficulties, given music's ubiquitous capacity to regulate emotional state (Saarikalio, 2009, 2013).

5.5.2 Conclusions

Overall, findings suggest divergence between audio conditions in terms of relationships with socio-emotional dimensions in children with difficulties. Although the lack of links between prosody recognition accuracy and externalising difficulties appears at odds with some past research (e.g., Chronaki et al., 2015a), it may be that differences in participant characteristics, particularly age, partially explain this finding. Findings also suggest an independent link between musical emotion recognition and externalising difficulties. Although correlational in nature, it may be that emotional understanding of music, as a specific aspect of musical development, partially explains links between broader musical abilities and externalising problems (e.g., Kim & Kim, 2018). Additionally, musical emotion recognition may to some extent mirror children's broader emotional abilities (e.g., Escoffier et al., 2013) and their links to externalising problems (e.g., Cooper et al., 2020), although the lack of corresponding association for vocal prosody questions this view. Irrespective of the

mechanisms involved, future longitudinal research would add clarity to the research base. Such research could inform music-based interventions for children with socioemotional, behavioural and cognitive difficulties - elucidating the specific role of emotional aspects of musical understanding in relation to these difficulties and pinpointing the most efficacious targets/approaches for intervention.

6. General Discussion

6.1 Overview and Aims

Both vocal prosody emotion recognition and a range of musical experiences/abilities are associated with positive socioemotional and behavioural adjustment in children (Blasco-Magraner et al., 2021; Chronaki et al., 2015a; Neves et al., 2021; Politimou et al., 2018). Concurrently, music and voice are closely related via their ability to elicit and convey emotion (Juslin & Sloboda, 2011), which has been linked to similarities in evolved expressive acoustic patterns between audio conditions (Juslin & Laukka, 2003; Scherer et al., 2015). However, there is disagreement regarding the extent to which expressive acoustic features align with discrete emotions, or more fundamental features arousal (low to high energy) and valence (negative to positive feeling - Cespedes-Guevara et al., 2018). Regardless of the nature of these expressive similarities, music and voice align in relation to the typical development of emotion recognition accuracy (Heaton & Allgood, 2015; Vidas et al., 2018), often attributed to shared mechanisms at the acoustic-perceptual level (e.g., Vigl et al., 2024). Given these expressive and perceptual links between musical and vocal emotions, and links between music and socio-emotional development more generally, music may be a powerful avenue through which to engage with and understand emotions, with possible intervention applications relating to socio-emotional development.

However, the few studies that have explored both children's musical and vocal emotion recognition have not directly considered factors contributing to observed similarities and differences between conditions. Understanding of these elements may be particularly important for research into individual differences in musical and vocal emotion recognition development. However, current research that has considered associations between vocal emotion recognition/musical abilities and broader socio-emotional dimensions are largely

based on typically developing samples. Reflecting this, proposals regarding a possible cross-condition model of audio emotion recognition development (Heaton & Allgood, 2015) lack clarity in terms of underpinning explanations for emotion recognition patterns, applicability beyond typical development, and broader implications for socio-emotional development. Further, despite the early significance of interpreting emotional information in singing for a range of developmental outcomes (Politimou et al., 2018; Schubert & McPherson, 2015), the development of emotion recognition for singing stimuli has received limited attention. Accordingly, drawing on multi-stage models of audio emotion recognition (Bestelmeyer et al., 2014; Schirmer & Kotz, 2006), this thesis aimed to begin to integrate these elements into a better explicated cross-condition model of audio emotion recognition development. This model could inform music-based interventions for children.

To meet this aim, the experimental chapters in this thesis answered the following overarching research questions:

Chapter 2: Do adults' perceptions of emotions, and valence and arousal, differ between musical and vocal stimuli, and how do patterns relate to stimulus acoustic features?

Chapter 3: In TD children, are emotion recognition patterns similar for musical and vocal stimuli, and how do emotion perceptions relate to stimulus-level arousal and valence dimensions?

Chapter 4: Are there individual differences in children's emotion recognition accuracy for musical and vocal stimuli, and can these be partially explained by general and/or emotion-specific language development?

Chapter 5: Is emotion recognition associated with broader socio-emotional dimensions in children with high levels of difficulties in these domains, and are these associations similar between musical and vocal conditions?

Given overlap between chapter findings, findings from chapters are integrated and discussed in relation to the following themes:

1. Similarities and differences between audio conditions in adults' and children's emotion recognition patterns.
2. Stimulus-level explanations for observed similarities and differences in emotion recognition patterns between audio conditions.
3. Emotion language and its possible role as a condition-general mechanism of audio emotion recognition development.
4. Associations between emotion recognition accuracy for each audio condition and socio-emotional dimensions.

Findings are then discussed in terms of their theoretical implications for the proposed cross-condition model of audio emotion recognition presented in Chapter 1 (section 1.6.3). Practical implications of the thesis findings are then discussed, before some strengths, limitations, and future directions are outlined.

6.2 Musical and Vocal Emotion Recognition Patterns

6.2.1 Audio Emotion Recognition – Broad Convergence but Condition-Specific Patterns

In support of hypothesised overall similarities in how musical and vocal emotions are perceived, the present thesis found positive correlations between emotion recognition accuracy for audio conditions, across experiments, aligning with past research with adults and TD children (Laukka & Juslin, 2007; Vidas et al., 2018). Further, there was alignment between audio conditions in typical development trajectories for overall recognition accuracy in Chapter 3 (particularly for instrumental music and prosody), in line with past research with TD children (Heaton & Allgood, 2015; Vidas et al., 2018). The steep developmental trajectories for emotion recognition accuracy in Chapter 3 also supported the presence of a

particularly important developmental period between 4-9 years in vocal prosody (Grosbras et al., 2018; Chronaki et al., 2015b), and instrumental music (Heaton & Allgood, 2015). Taken together, the current findings align with previous research to support a proposed cross-condition model of audio emotion recognition development (Heaton & Allgood, 2015). This aligns with the proposed role of music as a super-expressive voice - exaggerating acoustic information that aligns with specific emotions in voice and music (Juslin & Sloboda, 2011; Scherer, 1995). Importantly, correlations between conditions were particularly strong in samples of children with higher levels of variation in socio-emotional difficulties and emotion recognition accuracy (Chapters 4 and 5), even when accounting for age. This adds robustness to associations between emotion recognition accuracy for audio conditions and may support the possibility of shared underlying expressive and/or perceptual mechanisms for musical and vocal stimuli (Juslin & Laukka, 2003; Vigl et al., 2024).

However, findings in Chapters 2 and 3 also indicated some differences between audio conditions in adults' and TD children's recognition accuracy for certain emotions. This may question the extent to which music and voice are linked via similar expressive features for each emotion (Juslin & Laukka, 2003). For example, in both adults (Chapter 2) and TD children (Chapter 3), anger was easier to recognise in prosody, and happiness easier to recognise in instrumental music, relative to other conditions. This aligns with past research, indicating that anger is salient early in development for prosody stimuli (Sauter et al., 2013), but difficult to recognise in music (Vidas et al., 2018). Conversely, while happiness is recognised accurately in music from an early age (Dalla Bella et al., 2001; Vidas et al., 2018), general negative biases during prosody emotion recognition and specific difficulties recognising happiness can persist through adolescence and into adulthood (Chronaki et al., 2018; Lausen & Hammerschmidt, 2020). These findings question the idea that music represents a directly acoustically matched counterpart to the voice, grouped around emotion

categories (Juslin & Sloboda, 2011). This has implications for any cross-condition audio emotion recognition model, as condition-specific patterns need to be accounted for. The inclusion of singing stimuli offered possible explanation for some of these between-condition differences, while highlighting its importance to audio emotion recognition development.

6.2.2 Singing – Distinct but Developmentally Significant

Singing was included in the present thesis due to its position as a possible conceptual and developmental bridge between instrumental and vocal expressions. Conceptually, singing is tied to vocal prosody via their shared physiological constraints (Scherer et al., 2017), and to music via musical conventions such as musical mode (Balkwill & Thompson, 1999). Findings indicated that the ability to recognise emotion in singing stimuli aligned with the emotion recognition of prosody and instrumental music in different ways. Adults and children both with and without socio-emotional difficulties had more difficulty recognising sung emotions relative to both instrumental and prosody stimuli (Chapters 2, 3, 4, and 5). However, in Chapter 3, findings suggested that this difference may be less pronounced earlier in development. This aligned with past research, indicating an early ability to distinguish happiness and sadness in singing with comparable accuracy to instrumental music (Franco et al., 2017; Morton & Trehub, 2007). In addition, support for thesis findings comes from existing evidence that singing is more difficult than prosody to recognise for adults (Livingstone & Russo, 2018).

In terms of how prosody and instrumental music differed in their relation to singing stimuli, emotion recognition and confusion patterns for singing aligned more closely with prosody than instrumental music. For example, in Chapters 2 and 3, the salience of happiness in instrumental music did not translate to singing, despite a shared valence indicator in musical mode (Gomez & Danuser, 2007), while confusion between fear and sadness was unique to singing and prosody stimuli. This may suggest heightened emphasis on the

physiological constraints pertinent to the voice, and their close links to expressed arousal (Scherer et al., 2017), relative to valence-related music-specific expressive features such as mode (Gomez & Danuser, 2007).

Findings in Chapter 3 also indicated that children's emotion recognition of singing stimuli predicts accuracy in both instrumental and prosodic conditions, independent of age, music training, and accuracy for the other condition. As discussed in Chapter 3, the ability to draw meaning from 'song-like' early caregiver-child interactions, underpinned by developing sensitivity to acoustic features such as loudness, timing patterns, and pitch (Trehub, 2001; Flom & Bahrick, 2007), may form the basis of advancing emotion understanding across audio domains (Boone & Cunningham, 1998; Vanden Bosch der Nederlanden et al., 2022). Thus, despite greater difficulty recognising sung emotions in children both with and without socio-emotional difficulties, its inclusion adds clarity to a developmental model, both in terms of the emphasis placed on certain perceptual features, and developmental explanations for similarities and differences between conditions. Longitudinal research assessing children's early sensitivity to emotion in singing, and later instrumental/prosody emotion recognition, would strengthen inferences.

Despite differences in emotion recognition accuracy between conditions, at the level of adults' valence and arousal perceptions in Chapter 2, singing generally aligned with other audio conditions. Between-condition similarities were particularly pronounced for arousal - there was no evidence for differences between audio conditions in how adults perceived emotions in terms of arousal levels, while there were some slight differences in relation to perceived valence (happiness was perceived as more positive in instrumental stimuli relative to other audio conditions – Chapter 2). This may offer support for theories that centralise expressed valence, and especially arousal, within emotion recognition similarities between music and voice (Cespedes-Guevara et al., 2018). Consideration of stimulus acoustic features

in relation to perceptions of affective dimensions and emotions added clarity to the possible importance of arousal expression and comprehension for audio emotion recognition.

6.3 Stimulus-Level Mechanisms of Audio Emotion Recognition – Acoustic Correlates and Associations with Affective Dimensions

6.3.1 Acoustic Features and Their Links to Perceptual Patterns

Within this thesis, acoustic analyses should be treated with caution given the small set of stimuli and acoustic features, and the correlational approach to their analysis. However, findings in Chapter 2 indicated some between-condition similarities in associations between acoustic features and adults' perceptions of emotions (the proportion of time the given emotion was selected), valence, and arousal. Between-condition similarities were particularly prominent in relation to arousal. Indeed, acoustic features predicted adult's arousal ratings of stimuli with very similar strength across audio conditions, while correlations with specific features also aligned to some degree (loudness and tempo/speech rate across conditions, plus brightness for prosody and singing stimuli). For valence, the predictive strength of acoustic features was lower for all conditions relative to arousal, but higher in the instrumental and singing conditions relative to prosody, aligning with past research (Weninger et al., 2013). There were no associations between valence and specific acoustic features that aligned across all three audio conditions. These findings agree with past research suggesting similarities in the acoustic features expressing arousal across audio conditions, but differences for valence (Llie & Thompson, 2006; Scherer et al., 2015).

Acoustic feature levels were also analysed in relation to emotion perceptions in Chapter 2. There were few cross-condition correlations between acoustic features and emotion perceptions. Correlations that were consistent across conditions were mostly for features related to arousal – loudness, tempo/speech rate, and brightness. Although other

features, such as pitch, were related to the perception of some emotions, the direction of these associations and specific emotions involved varied between conditions. The lack of cross-condition correlations between acoustic features and emotion perceptions does not align with theories suggesting cross-condition mapping of acoustic features to specific emotion categories (Juslin & Laukka, 2003). As discussed, differences in adults' and children's emotion recognition accuracy for certain emotions also questions this premise (Chapters 2 and 3). Instead, the cross-condition salience of arousal-related features such as loudness and tempo/speech-rate (Chapter 2) may support theories proposing that the closest expressive links between music and voice relate to arousal (Cespedes-Guevara & Eerola, 2018).

Although correlations between specific acoustic features and emotion perceptions differed between audio conditions, the overall predictive strength of the set of acoustic features on adults' emotion perceptions was generally high, and of similar strength between instrumental and prosody conditions (Chapter 2). This supports research stressing that similar acoustic features are important for emotion communication in music and voice (Eerola et al., 2013; Juslin & Laukka, 2003). Considered alongside observed correlations between overall emotion recognition accuracy for audio conditions in adults and children (Chapters 2, 3, 4, and 5), but between-condition differences in accuracy for certain emotions (Chapters 2 and 3), this may support the idea of condition-general acoustic processing mechanisms, without direct cross-condition mapping of acoustic features to emotions. For example, individual differences in musical aptitude – the ability to detect variations in acoustic features – has been evidenced as a cross-condition predictor of emotion recognition accuracy, over and above any formal music training (Jansen et al., 2023; Vigl et al., 2024). Indeed, the present thesis found no evidence for a link between years of formal music training and emotion recognition accuracy once age was accounted for (although the overall low levels of music training in Chapters 3-5 limit the strength of these inferences). It may be that improvement in this

general ability to perceive acoustic variations is one avenue through which musical interventions can have diverse positive effects on children's socio-emotional development, including their vocal emotion recognition accuracy (Blasco-Magraner et al., 2021). An explanation of emotion recognition similarities between music and voice via musical aptitude also aligns with theories of the evolutionary basis of musical and vocal emotions. For example, some researchers claim that prosodic aspects of early communication formed the basis of parallel, but independent, evolution of emotion communication through music, vocalisations, and language (Brown, 2017; Clark et al., 2015). Seemingly, future research directly examining individual differences in musical aptitude in relation to musical and vocal emotion recognition development could have important implications, both theoretically and practically.

Acoustic patterns for singing stimuli also supported possible cross-condition salience of arousal expression for audio stimuli, and added clarity to the developmental role singing might play within audio emotion recognition development. As discussed (section 6.3.2), the ability to recognise emotions in singing stimuli predicted accuracy for both prosody and instrumental conditions in TD children (Chapter 3), in line with suggestions that interpretation of emotional information in early song-like expressions forms the basis of later developing emotion recognition ability across audio domains (Boone & Cunningham, 1998; Vanden Bosch der Nederlanden et al., 2022). However, acoustic features generally predicted perceptions of emotions less strongly for singing stimuli relative to instrumental and prosody conditions, reflected in overall difficulties recognising sung emotions (Chapters 2, 3, 4, and 5). Similarities between singing and other audio conditions were more pronounced in relation to arousal – there was no between-condition difference in adults' arousal perceptions for each emotion, or in how strongly arousal perceptions were predicted by acoustic features (Chapter 2). Accordingly, it may be that early singing-based caregiver-child interactions represent a

context through which early arousal-related comprehension develops. Indeed, the exaggeration of acoustic qualities such as pitch, timing, and loudness within these interactions (Nakata & Trehub, 2011; Trehub & Trainor, 2011), coupled with their function as modulators of arousal in both infant (Corbel et al., 2016; Nakata & Trehub, 2004) and caregiver (Cirelli et al., 2020), make them a key context through which to develop early associations between acoustic-features and expressed arousal. As is discussed below, this arousal-based understanding may be key to audio emotion recognition development.

6.3.2 Arousal-Based Broad-to-Differentiated Emotion Recognition Development

Widen (2013) posits that emotion recognition development involves age-related increase in understanding of fundamental dimensions valence and arousal, before development of more differentiated understanding of specific emotion categories (the 'broad-to-differentiated' model of emotion recognition). For facial stimuli, valence appears particularly important to emotion recognition development (Woodward et al., 2022). Understanding of arousal and valence are also integral to audio emotion recognition, with ratings for these dimensions strongly predictive of emotion recognition patterns for both music and voice (Eerola & Vuoskoski, 2011; Sauter et al., 2010). However, development of valence comprehension for vocal stimuli may lag behind that for facial expressions and continue beyond 5-years (Nelson & Russell, 2011). It may also be the case that arousal plays a stronger role in the perception of audio emotions relative to facial expressions, perhaps due to strong links between perceived arousal and acoustic features (Chapter 2; Cespedes-Guevara & Eerola, 2018; Weninger et al., 2013).

In Chapter 3, to assess the possible roles of arousal and valence understanding within audio emotion recognition development, normative levels of arousal and valence were assigned to each audio stimulus based on adults' average ratings from Chapter 2, in line with past research (e.g., Woodward et al., 2022). The extent to which stimulus arousal and valence

predicted TD children's emotion perceptions was then assessed, to examine any similarities and differences between conditions, and changes with age, in how these fundamental affective dimensions relate to emotion perceptions. Findings offered some support for the salience of expressed arousal during children's emotion perceptions across audio conditions. Although there were some differences between conditions in how stimulus arousal related to emotion perceptions, this was explained by differences in effect strength, which was generally strong across conditions. Indeed, in general, the strength of associations between stimulus arousal and children's emotion perceptions were stronger than those for valence, across audio conditions (Chapter 3). Further, the most prominent cross-condition confusion pattern in both TD children and adults was between sadness and calmness – two emotions similar in arousal, but not valence (Chapters 2 and 3). This aligns with children's emotion recognition confusion patterns between emotions similar in expressed arousal in previous prosody (Nelson & Russell, 2011; Grosbras et al., 2018), instrumental (Kragness et al., 2021; Dalla-Bella et al., 2001), and singing research (Adachi & Trehub, 1998). This arousal salience may relate to early sensitivity to features such as tempo/speech rate and loudness. Indeed, children as young as four are sensitive to these acoustic properties and can draw on them to discriminate emotions in instrumental and singing stimuli (Adachi & Trehub, 1998; Kragness et al., 2021; Mote et al., 2011). Thus, findings offer some support for the suggestion that arousal is the most salient affective dimension underpinning audio emotion recognition, contrary to the apparent importance of valence for facial emotion recognition (Cespedes-Guevara & Eerola, 2018; Widen, 2013; Woodward et al., 2022).

Findings in Chapter 3 suggested that the relationship between stimulus valence and emotion perceptions increased with age - happiness and calmness selections became more common for stimuli more positive in valence, while anger, fear, and sadness selection became more common for stimuli more negative in valence. This aligns with past research for

musical and vocal stimuli suggesting increasing understanding of expressed valence with age (Kragness et al., 2021; Nelson & Russell, 2011). However, its relationship with emotion perceptions also differed between conditions, more prominently than was the case for stimulus arousal. Indeed, in some cases, the relationship between valence and emotion perceptions differed in *direction* between conditions, suggesting pronounced differences in how valence related to emotion perceptions. For example, for prosody stimuli only, erroneous selections of anger and fear were related to increasing (more positive) stimulus valence, reflected in confusion between these negative emotions and happiness (Chapter 3). This between-valence confusion pattern appeared to persist with age for prosody stimuli. However, for instrumental music, this confusion pattern was apparent only in young children (4-5-years), and negligible by 6-7-years (Chapter 3). Reduction in this confusion pattern for instrumental music may relate to developing understanding of musical mode – strongly related to perceived valence and thought to develop around 6-8-years (Gomez & Danuser, 2007; Dalla-Bella et al., 2001). Accordingly, it may be that some music-specific features lead to differences between conditions in how valence relates to emotion perceptions across development.

As discussed in Chapters 2 and 3, findings may suggest arousal is a salient cross-condition communicator of emotional information, with valence and emotion concept knowledge acting as 'affective semantic indicators' that refine arousal-based perceptions (Holz et al., 2021). Given age-related change in relationships between stimulus valence and emotion perceptions and reductions in between-valence confusion patterns (Chapter 3), it may be that this valence and emotion category understanding develops later, facilitating more differentiated recognition of audio emotions. This may suggest that the broad-to-differentiated model of emotion recognition, when applied to audio stimuli, requires greater emphasis on comprehension of expressed arousal, especially early in development when

conceptual understanding of emotions and valence-based understanding of audio stimuli remains limited (Nelson & Russell, 2011; Kragness et al., 2021; Nook et al., 2020). However, there may be differences between conditions in how this valence understanding develops, given differences in the expressive cues available.

6.4 Emotion Language as a Condition-General Mechanism of Audio Emotion Recognition Development

Audio emotion recognition involves multiple stages, including the low-level perception and integration of acoustic input, and the cognitive interpretation of this input as an emotion category (Bestelmeyer et al., 2014). Within this cognitive stage of emotion recognition, Chapter 4 indicated that children's general verbal ability was correlated with emotion recognition accuracy across audio conditions. However, when analysed together, emotion language comprehension was the only significant independent linguistic predictor of audio emotion recognition accuracy. This association appeared robust - predicting emotion recognition accuracy for all conditions, across the whole sample and within the referred/TD samples in isolation. Accordingly, emotion language comprehension could be a condition-general mechanism that partially explains the significant correlations in children's emotion recognition accuracy between musical and vocal conditions in the present thesis (Chapters 3, 4, and 5). Indeed, evidence suggests that musical and vocal emotion recognition relies on condition-general mechanisms for social cognition beyond similarities between conditions at the acoustic-perceptual level (Escoffier et al., 2013; Schirmer et al., 2012; Lima et al., 2016).

The unique association between emotion language comprehension and audio emotion recognition accuracy aligns with the theorised importance of emotion-specific language development for organising emotion knowledge within emotion categories, facilitating emotion recognition development (Lindquist, 2017). Indeed, emotion language comprehension is associated with facial emotion recognition accuracy across childhood

development (Beck et al., 2012; Streubel et al., 2020). The presence of a similar association between emotion language comprehension and musical emotion recognition (Chapter 4) aligns with past research indicating that children's expressive emotion language fluency predicted individual differences in instrumental emotion recognition accuracy, while a measure of general verbal fluency did not (Plate et al., 2022). The present findings suggest that this specific link also exists for *receptive* emotion language comprehension, and for other types of musical stimuli, such as singing. The present thesis also extends past research by demonstrating a unique association between emotion language comprehension and vocal emotion recognition in children. Accordingly, emotion language comprehension may mediate the developmental associations between general verbal ability and vocal emotion recognition accuracy found in past research (Griffiths et al., 2020), in line with research with facial expressions (Streubel et al., 2020). However, general language abilities may also have an independent link with emotion recognition during some stages of development (Grosse & Streubel, 2024). More research including a broader age range, and children both with and without socio-emotional difficulties, could help clarify the specific roles of general and emotion-specific language abilities during the development of audio emotion recognition.

It is important to note that the directional association between emotion-specific language development and emotion recognition, whereby language ability precedes emotion recognition, while theoretically sound, has not been demonstrated directly. While longitudinal research indicates a directional association between general language ability and vocal emotion recognition (Griffiths et al., 2020), the directional association between emotion language comprehension and audio emotion recognition proposed in Chapter 4 requires further research. In the case of music, especially, bidirectionality in this association is possible, due to apparent associations between the richness of children's home musical environment and their subsequent language development (Politimou et al., 2021), and

between various musical abilities and a range of later-developing language skills in pre-schoolers and children (Politimou et al., 2018; Papadimitrou et al., 2021). Although links between musical environment and emotion language understanding have not been demonstrated, nor links between early musical environment and emotional understanding of music, a bidirectional developmental link between emotion language and musical emotion recognition is plausible.

6.5 Between-Condition Differences in Associations with Socio-Emotional Dimensions

Chapter 5 examined associations between emotion recognition for audio conditions and internalising and externalising problems in children with generally elevated levels of difficulties in these domains. Findings provided evidence against an association between externalising and internalising difficulties and vocal prosody recognition accuracy, differing from some past research (e.g., Chronaki et al., 2015a), but aligning with others (e.g., Neves et al., 2021). As discussed in Chapter 5, certain methodological factors such as employing stimuli of differing intensities and adopting a younger sample may have contributed to this discrepancy (Chronaki et al., 2015a).

Conversely, Chapter 5 did find a significant negative association between instrumental emotion recognition accuracy and externalising difficulties. This association remained when statistically controlling for prosody emotion recognition accuracy, socio-demographic variables, and music training. These findings support the possibility of an association between musical emotion recognition and externalising difficulties, independent of any relationship with or through vocal emotion recognition ability. Past research has indicated associations between various musical abilities/interventions and fewer externalising problems in children (Gold et al., 2004; Kim & Kim, 2018; Rose et al., 2015). Further, associations have been identified between TD adolescents' ability to express emotions in music and their level of externalising difficulties (Saarikallio et al., 2014). Accordingly, the

present findings suggest that emotional understanding of music may be one avenue through which musical abilities/interventions relate to externalising problems in children, and that this association is also apparent in children with high overall levels of socio-emotional difficulties.

While the present thesis, and a range of past research, supports the presence of general cognitive mechanisms underpinning musical emotion recognition beyond acoustic-perceptual links to the voice (Chapter 4; Escoffier et al., 2013; Lima et al., 2016), reference to these mechanisms does not highlight anything specific to music that may underpin the observed association with the externalising dimension (Chapter 5). These music-specific links could relate to music-specific processing difficulties, or as discussed above, musical environment. The association found between musical emotion recognition and externalising difficulties could be related to children's musical environment, particularly early in development. As discussed in Chapter 5, the richness of children's musical environment, particularly exposure to infant-directed singing, could have a range of positive developmental implications, including for attachment with caregivers and emotion regulation (Politimou et al., 2018; Sanfillipo et al., 2021; Williams et al., 2015). Given the importance of factors such as attachment and emotion regulation to the development of externalising difficulties (Achenbach et al., 2016; Mullins & Henshaw, 2007), children's early musical environment could have important implications in relation to the externalising dimension. It could be hypothesised that early musical environment could relate to the presence of externalising difficulties through the mediation of musical emotion recognition ability. Given the exaggerated emotional expression pertinent to early music-like interactions with children (Trehub & Trainor, 2011; Nakata & Trehub, 2011), such a link is plausible. Overall, future longitudinal research considering this emotional aspect of musical development in relation to socioemotional dimensions would add clarity to the research base.

Another possible explanation for the specific association between instrumental emotion recognition and externalising difficulties in the present thesis relates to music-specific processing difficulties. For example, strong links between music processing and the reward system have been demonstrated (Blood & Zatorre, 2001; Zatorre & Salimpoor, 2013), while the reward system also modulates perceptions of musical emotions such that positive emotions are perceived as more pleasant, and some negative emotions as less pleasant, in adults with high reward-sensitivity to music (Fuentes-Sánchez et al., 2023). Atypicality in reward processing is also strongly related to externalising difficulties in children, with children with lower reward sensitivity often experiencing more externalising difficulties (Gatzke-Kopp et al., 2009; Kasparek et al., 2020). Plausibly, then, reward sensitivity could relate to both instrumental recognition accuracy and externalising difficulties, partially explaining the music-specific negative association with externalising difficulties within Chapter 5.

Chapter 5 findings suggest that despite associations between overall emotion recognition accuracy for audio conditions (Chapters 2, 3, 4, and 5), underpinned by condition-general mechanisms at the levels of acoustic perception/integration (e.g., Vigl et al., 2024), and cognitive interpretation (emotion language in Chapter 4), condition-specific associations with broader socio-emotional dimensions are possible. This has important implications for the proposed cross-condition model of audio emotion recognition.

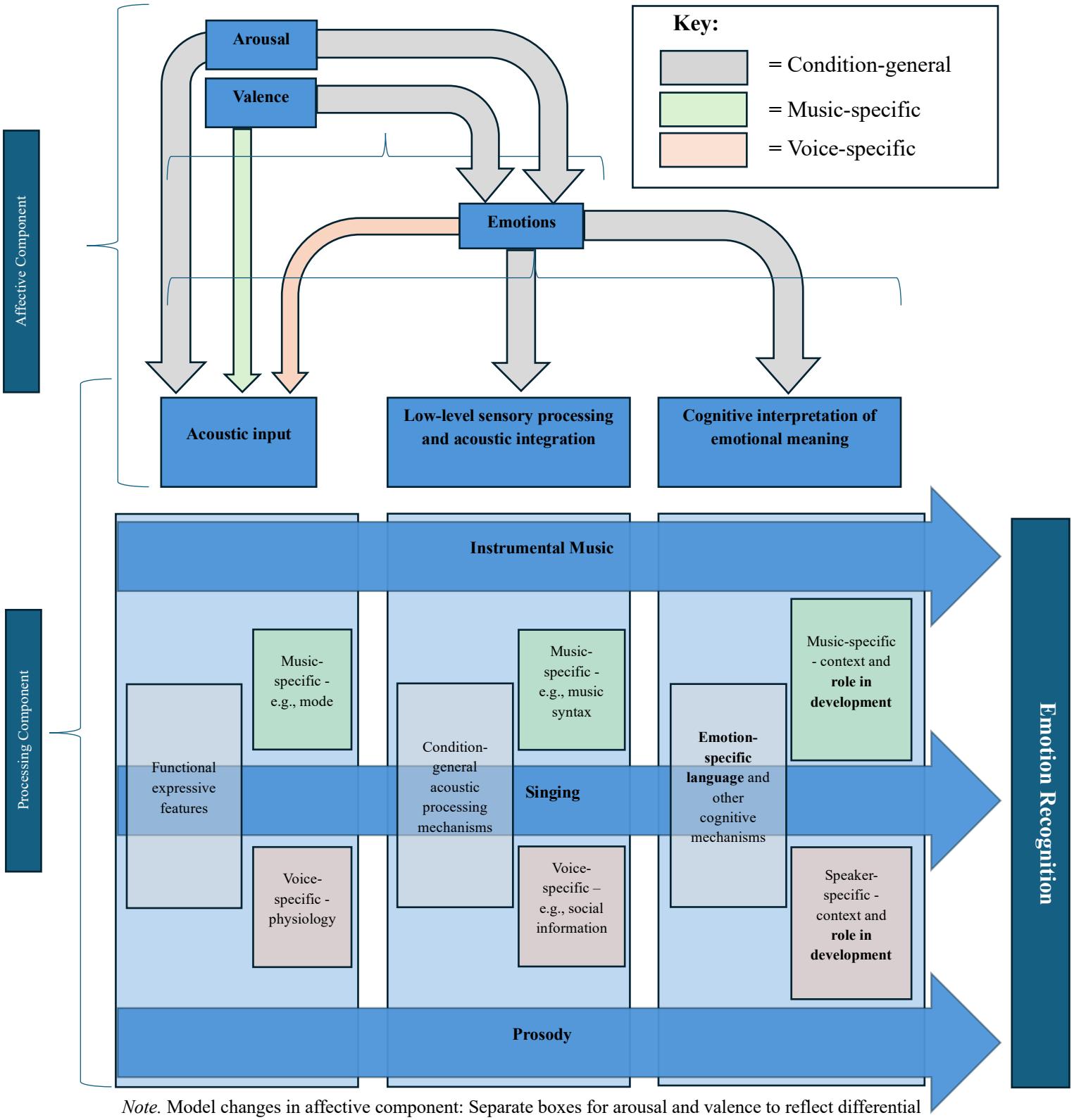
6.6 Thesis Implications

6.6.1 Theoretical Implications

6.6.1.1 Updating the Cross-Condition Model of Audio Emotion Recognition by Centralising Arousal and Accommodating Complexity.

Based on the findings of this thesis, a new cross-condition model of audio emotion recognition is presented in Figure 6.1 below. This model is an extension of Bestelmeyer et al.'s (2014) multi-stage model of audio emotion recognition (see figure 1.5 in Chapter 1). The original model involves three stages - acoustic input, acoustic perception/integration, and cognitive interpretation. There are also two 'components' of the model – the 'affective component' and the 'processing component'. The affective component denotes the independent and interactive role of arousal, valence and emotion categories, during emotion recognition, while the processing component outlines possible condition-general and condition-specific aspects involved in the processing of arousal, valance and emotion categories (for a more in-depth description of the model see Chapter 1, section 1.6.3). Based on findings in this thesis in relation to past research, three changes to the original model are discussed below.

Figure 6.1 - Updated Cross-Condition Model of Audio Emotion Recognition



Note. Model changes in affective component: Separate boxes for arousal and valence to reflect differential relationships with other parts of the model; Differing arrow colours to reflect condition-specific pathways within the affective component of the model. Model changes in processing component: Changes within the cognitive interpretation of emotion stage are marked in **bold.** Music-specific elements relevant to instrumental and singing conditions; voice-specific elements relevant to prosody and singing conditions.

A first adjustment to the original model (Figure 1.5) relates to the affective component of the model. Some dimensional theories posit that emotions are built from a two-factor understanding of arousal and valence (see Barrett & Bliss-Moreau, 2009). However, recent research has questioned this, with facial expression research indicating a more prominent role for valence (Woodward et al., 2022), in line with valence-focused theories of emotion recognition development (Widen, 2013). However, as discussed, findings in the present thesis point to a more prominent role for arousal in audio stimuli, especially early in development. For example, arousal may play a more prominent role in the acoustic input stage of the model due to its closer links compared to valence to expressive acoustic features (Bänziger et al., 2015; Chapter 2; Weninger et al., 2013). In Figure 6.1, this is reflected in the larger arrow between arousal and acoustic input. As discussed in Chapter 3, arousal may also play a relatively more important role in constituting children's emotion perceptions (Holz et al., 2021), particularly early in development. In Figure 6.1, this is reflected in the inclusion of different arousal and valence boxes, with independent arrows linking them to the emotion box. Although each arrow is equally weighted, it may be that the arousal arrow would be larger than the valence arrow early in development, while the valence pathway increases in strength with age (Chapter 3; Nelson & Russell, 2011).

There is however additional complexity to the affective component in relation to the relative prominence of valence between conditions, and its possible role during emotion recognition. For example, adults' valence perceptions were more strongly predicted by acoustic features in Chapter 2, and stimulus valence was more consistently related to children's emotion perceptions in Chapter 3, for musical relative to vocal stimuli. This was reflected in more between-valence emotion recognition confusion for vocal stimuli (Chapter 3). It may be that music-specific conventions such as musical mode and their close links to perceived valence led to these condition-specific patterns (Gomez & Danuser, 2007).

Therefore, although strong emphasis on arousal seems important for any cross-condition model (Cespedes-Guevara & Eerola, 2018), music-specific components at the acoustic input stage, and their specific ability to express valence, should also be emphasised. In Figure 6.1, this is reflected in the green music-specific arrow between valence and the acoustic-input stage of the model.

A final change in the affective component relates to possible direct communication of some emotions in the voice independent of arousal and valence. Indeed, past research has found that while acoustic features within vocal expressions are closely tied to perceived arousal, there are also 'unique combinations of distal cues and proximal percepts carrying information about specific emotion families, independent of arousal', particularly for emotions such as anger (Bänziger et al., 2015, para. 1). Such independent cues may have partially underpinned the high recognition accuracy for anger in children and adults in the present thesis (Chapters 2 and 3), despite anger being perceived similarly on arousal and valence in adults across conditions (Chapter 2), and the similar level of predictive strength of the current set of acoustic features on adults' anger perceptions across conditions (Chapter 2). As discussed in Chapter 3, the adaptive significance of understanding and responding to vocal anger signals from early in development may give rise to direct communication of vocal anger independent of arousal and valence dimensions (Buss, 2005). This is reflected in the red voice-specific arrow linking the emotion box to the acoustic input stage of the model in Figure 6.1.

Within the processing component of Figure 6.1, the current findings also point to the need to consider the role of developmental context. This is particularly salient in relation to singing stimuli. Indeed, Chapters 2 and 3 suggested singing is more closely matched to instrumental and prosody stimuli in terms of arousal and valence, rather than emotion categories. As discussed, this may relate to the developmental role of singing as a pre-

linguistic modulator of affective state (Nakata & Trehub, 2004; Corbel et al., 2016; Cirelli et al., 2020). Infant-directed song is also key in increasing attention towards the caregiver (Theissen & Saffran, 2009; Shön et al., 2008), while entraining them to broader sources of socio-emotional information (Lense et al., 2022). These unique developmental roles for singing may underpin not only the way it is perceived emotionally, but associations between the ability to recognise emotions in singing and broader socio-emotional dimension – as reflected in the singing-specific association with attention difficulties in Chapter 5. In Figure 6.1, the influence developmental context may have on how emotions are interpreted is highlighted within the cognitive interpretation stage of the model.

Finally, in relation to the cognitive interpretation stage of the model, findings from Chapter 4 supported the presence of a condition-general mechanism in emotion language comprehension, that operated independent of general language ability. This is reflected in a change in the cognitive interpretation stage of Figure 6.1 – changing 'language' to 'emotion-specific language' as a key condition-general cognitive predictor of audio emotion recognition accuracy. However, recent research with visual stimuli has suggested that emotion language comprehension may also play a role within earlier stages involving low-level perception and integration of stimuli (Brooks & Freeman, 2018; Lindquist, 2017). This would suggest a shift away from a consideration of dissociated acoustic integration and cognitive interpretation stages, as in Figure 6.1, towards a model that outlines how these processes dynamically interact during emotion recognition. An exploratory analysis assessed this possibility and indicated a possible interaction between individuals' emotion language ability and the strength of associations between acoustic features and emotion perceptions (Appendix T). However, the methods adopted were not well-suited to drawing any firm conclusions, as it was unable to examine whether higher comprehension of specific emotions related to the way in which that specific emotion was perceived. Accordingly, future research

could examine possible interaction between model stages in more detail via emotion word priming/impeding paradigms, which assess the influence of access to conceptual information (emotion words) on individual perceptions (Doyle & Lindquist, 2018; Nook et al., 2015; Satpute et al., 2016). This may lead to further model changes regarding interaction between model stages.

6.6.1.2 Indirect Implications for Discrete and Dimensional Theories of Audio Emotion Expression.

Although not a core aim of the current thesis, findings also have some indirect implications for debates between dimensional and discrete theories of vocal and musical emotion expression/perception (see Warrenburg, 2019). Discrete theories suggest expressive acoustic features align with emotion categories in a consistent way between conditions, predicating similarities in emotion recognition patterns/development (Juslin & Sloboda, 2011; Vidas et al., 2018). Conversely, dimensional theories claim that any commonalities between music and voice relate to similar expression of valence and arousal, which are categorised as emotions via language-mediated conceptual understanding (Cespedes-Guevara & Eerola, 2018). However, interplay between perceptual understandings of valence and arousal, emotion categories, and language, across development, suggests that emotions are best understood through an integrated approach that combines both discrete and dimensional perspectives (Ruba et al., 2018; Widen & Russell, 2002; Widen 2008; Woodward et al., 2022 – see Chapter 1, section 1.3.5 for discussion). The current findings appear to align most closely with this integrated theoretical approach. Indeed, condition-specific emotion recognition patterns (Chapters 2 and 3) and correlations between emotion perceptions and acoustic features (Chapter 2) argue against the close cross-condition alignment of expressive features to emotion categories within discrete theories (Juslin & Laukka, 2003). Conversely, the greater salience of arousal compared to valence in audio stimuli (Chapter 2; Chapter 3;

Holz et al., 2021), and certain possible direct links between expressive features and some vocal emotions (Bänziger et al., 2015) does not align with the assumption that arousal and valence are equivalent building blocks of emotion for audio stimuli (Barrett & Bliss-Moreau, 2009). Seemingly, findings support a bridging of discrete and dimensional theories, via consideration of expressive perceptual features that align with both emotion categories and affective dimensions, as well as more conceptual processes relating to emotion language (Woodward et al., 2022).

6.6.2 Practical Implications

Findings from this thesis also have a range of practical implications. As described in Chapter 1, musical abilities and training have been linked to a range of developmental outcomes, including those related to the externalising dimension (Boucher et al., 2021; Kim & Kim, 2018; Rose et al., 2015). Given the correlations between instrumental emotion recognition accuracy and externalising problems observed in Chapter 5, it may be that emotional understanding of music, specifically, is one avenue through which these musical interventions/abilities relate to externalising difficulties. As discussed in Chapter 5, it may be that emotion recognition for music, particularly instrumental music, reflects ability in relation to acoustic-perceptual mechanisms (Vigl et al., 2024), and more general socio-emotional processing mechanisms (Escoffier et al., 2013; Lima et al., 2016). Accordingly, an intervention focus on understanding emotions in instrumental music may be an effective avenue for intervention, particularly for children with externalising difficulties. However, given the correlational nature of this finding, future longitudinal or intervention research would strengthen this inference. Further, past research has found links between children's musical environment and a range of positive developmental outcomes (Politimou et al., 2018), and some important predictors of externalising problems in children (Williams et al., 2015). Accordingly, future research could examine the interrelations between children's home

musical environment, their developing emotion understanding of music, and externalising difficulties. A longitudinal focus on these elements could elucidate the nature and directionality of specific developmental links between musical abilities and externalising problems and identify key developmental stages for intervention.

The present thesis also aligns with research stressing the developmental importance of singing (Schubert & MacPherson, 2015). This may facilitate singing-based interventions with broader socio-emotional outcomes as targets. For example, given that singing emotion recognition accuracy predicted recognition accuracy in other audio domains, interventions encouraging singing with children may have positive implications for the development of emotion recognition in other audio conditions. Given associations between instrumental recognition accuracy and externalising difficulties in the present thesis (Chapter 5), and associations between vocal emotion recognition and other dimensions such as behavioural and cognitive self-regulation found in past research (Neves et al., 2021), these singing-based interventions could have diverse indirect positive effects. Interventions may be particularly efficacious early in development, while understanding of arousal-based expressive cues rapidly expands, as does their ability to regulate their own levels of arousal (Flom & Bahrick, 2007; Shenfield et al., 2003; Trehub, 2001). Singing-based interventions also require less equipment/training, making them more widely accessible and increasing the possibility of further direct benefits (e.g., for caregiver/child attachment - Lense et al., 2022; Sanfillipo et al., 2021).

The current findings also highlight possible music-based intervention approaches that aim to highlight the ways in which audio stimuli, including vocal stimuli, expresses emotion. Indeed, past research indicated a positive impact of music-based interventions on vocal emotion recognition accuracy, including those focused specifically on emotion expression in music (Mualem & Lavidor, 2015). Given the proposed developmental importance of

expressed and perceived arousal in the present thesis (Chapters 2 and 3) and past research (e.g., Bänziger et al., 2015; Cespedes-Guevara & Eerola, 2018; Holz et al., 2021), future interventions focussed on improving vocal emotion recognition could focus on this affective feature. For example, an intervention could involve focus on certain cross-condition arousal-based cues, such as loudness and tempo, and demonstrate how these are varied in both music and voice to express arousal and emotions. The ability to manipulate music in real-time may make music a powerful avenue through which to explore these arousal-based expressive variations (Kragness et al., 2021), More generally, music-based interventions can benefit from positive social benefits of group music making (Linnavalli et al., 2021; Mullensiefen & Harrison, 2020).

Finally, given the apparent condition-general importance of emotion language comprehension to emotion recognition development (Chapter 4), intervention approaches could incorporate a focus on this ability. Given the capacity of music to convey socially relevant semantic information (Fritz et al., 2019; Koelsch et al., 2004), music may offer a powerful tool through which to encourage children to build the perceptual and conceptual skills necessary for emotion recognition (Nenchevca et al., 2023; Woodward et al., 2022) However, more research is needed regarding the interplay between these language-mediated conceptual processes, and low-level acoustic perception, to facilitate interventions focused on specific developmental mechanisms.

6.7 Strengths, Limitations and Future Directions

6.7.1 Strengths

One strength of the current thesis was its focus on various levels of explanation for emotion recognition. Consideration of stimulus-level features, including acoustic features and affective dimensions arousal and valence, strengthened inferences regarding similarities and

differences between musical and vocal emotions. For example, consideration of stimulus level features highlighted similarities between conditions in relation to the expression and perception of arousal, but some more pronounced differences in relation to valence. This helped to explain certain condition-specific recognition patterns, such as the relative salience of anger in vocal expressions and happiness in instrumental music (Chapters 2 and 3). At the cognitive level, consideration of emotion language comprehension highlighted a condition-general mechanism that could partially account for overall similarities in emotion recognition accuracy development, despite these stimulus-level discrepancies (Chapter 4). As discussed, better understanding both the similarities and differences between conditions could have important implications for future research and interventions.

A second strength relates to the inclusion of a diverse range of children. Audio emotion recognition research with children beyond typical development is currently very limited, and generally isolated to specific diagnoses (e.g., autistic children – Sivathasan et al., 2023). The transdiagnostic approach adopted here aligned with the focus on underlying mechanisms of emotion recognition development and ensured that findings were reflective of the varying levels of difficulties experienced by the population of focus (Astle et al., 2022). Further, findings indicated that correlations between recognition accuracy for audio conditions were marginally stronger within groups of children with higher overall levels of socio-emotional difficulties, strengthening assertions regarding shared expressive and developmental mechanisms.

A final strength relates to the analytical approach adopted, including the use of mixed models and the use of BFs alongside frequentist statistics. Mixed models are favoured over fixed effects models or ANOVA-based methods as they avoid the need for data aggregation. Aggregation can lose information about variability – such as within individuals or within items (or groups of items, such as emotion categories) – which reduces statistical power and

increases the chance of false positives (Barr et al., 2013). Accordingly, the findings described in the thesis are more robust and generalisable than they would have been had more traditional models been used. The present thesis also integrated Bayesian findings, in the form of BFs. BFs are able to provide direct evidence for the strength of both the alternative and null hypotheses (Keysers et al., 2020). In the present case, this was important, as many predictions and inferences related to a lack of differences between conditions. For example, similarities in overall recognition accuracy for instrumental and prosody stimuli in Chapters 2 and 3 could be interpreted as a genuine similarity, rather than an absence of sufficient evidence for a difference.

6.7.2 Limitations and Future Directions

There are also limitations in this current body of work. First, although consideration of stimulus-level acoustic features represented a strength, the inferences drawn are speculative, given the small set of stimuli and acoustic features, and the correlational approach to analysis. The relatively limited number of trials per condition was dictated by the decision to include three types of audio condition and by concern for maintaining the attention of the child participants. Research with a more expanded stimulus set would strengthen inferences regarding associations between acoustic features and emotion perceptions. The acoustic features of the stimuli were selected based on past cross-condition research (e.g., Llie & Thompson, 2006; Coutinho et al., 2013; Paquette et al., 2018; Jusin & Laukka, 2003) and reduced based on collinearity between features. Future cross-condition research could examine a wider set of acoustic features in relation to developmental perceptual patterns, as this may provide a fuller picture regarding cross-condition similarities and differences at the acoustic-perceptual level. In addition, the acoustic analyses in the current thesis did not speak directly to acoustic sensitivities. Given the suggested importance of general mechanisms at the acoustic-perceptual level for musical and vocal emotion

recognition development (Vigl et al., 2024), establishing acoustic sensitivities via psychoacoustic thresholds could allow for more in-depth understanding of how comprehension of specific acoustic features relates to emotion recognition, both in adults and across development.

A second limitation relates to the use of piano music as instrumental stimuli. This decision was based on the availability of validated stimuli that covered a sufficient number of emotion categories to allow detailed comparison with vocal stimuli. Although stringed stimulus sets exist (e.g., Paquette et al., 2013), they tend to only include two or three emotion categories. However, the type of instrumentation can affect emotion perceptions, independent of structural musical features such as loudness, tempo, and melody (Hailstone et al., 2009). Further, it may be that stringed instruments are better able to approximate certain expressive vocal devices, such as vibrato and its likeness to vocal tremor (Bedoya et al., 2021). As such, future research could validate and adopt musical stimuli that encompasses other types of musical instrument. Alternatively, music stimuli including multiple instruments with the aim of maximally expressing each emotion could be revealing regarding the extent to which music can communicate emotions to children and adults.

The findings are also limited by the use of a cross-sectional sample to explore developmental mechanisms. The findings require corroboration via more robust longitudinal methods. For example, inferences regarding the possible developmental significance of emotion understanding of singing stimuli (Chapter 3) would be strengthened via longitudinal research. Similarly, in line with the approach of Griffiths et al. (2020) in relation to general verbal ability, claims implicating emotion language comprehension as a possible condition-general developmental mechanism of audio emotion recognition (Chapter 4) would also be strengthened via longitudinal research. As noted, although the directionality of this association is based on theory (Lindquist, 2017), possible bidirectionality in links between

musical and linguistic development (Politimou et al., 2018) calls for further examination.

Finally, the present findings and inferences should not be over-generalised, particularly across different cultures. This is because culture will likely modulate processing within different stages of the cross-condition emotion recognition model discussed (Figure 6.1). For example, at the acoustic input and integration stages, development of some musical abilities, including rhythm discrimination, are strongly influenced by culture (Stewart & Walsh, 2005), while certain expressive acoustic features are utilised and perceived differently across cultures (Athansopooulos et al., 2021; Laukka et al., 2013a). At the cognitive interpretation stage, the integral role for emotion language comprehension in audio emotion recognition evidenced in the present thesis indicates that cross-cultural variation in how emotions are defined and understood may strongly influence variation in emotion recognition patterns (Lindquist et al., 2022). Collectively, this suggests that the present findings may be to some degree specific to Western musical conventions and conceptualisations of emotion. Future cross-cultural research examining emotion recognition development across audio modalities, and possible stimulus-level and cognitive mechanisms, could be informative in this regard.

6.8 Conclusions

Extensive theoretical and empirical research posits an evolved adaptive, functional link between musical and vocal emotions (Juslin, 2018; Brown, 2017; Clark et al., 2015), which translates to between-condition similarities in emotion recognition development (Vidas et al., 2018; Heaton & Allgood, 2015). However, from a developmental perspective, understanding of some of the mechanisms underpinning similarities and differences between musical and vocal emotion recognition, as well as associations between these factors and broader socio-emotional dimensions, remains limited. Further, individual differences in vocal and musical emotion recognition development have been underexplored - restricting the

potential of evidence-based musical interventions/practices, particularly for children with socio-emotional difficulties. Accordingly, this thesis aimed to extend an existing cross-condition model of audio emotion recognition development (Heaton & Allgood, 2015), adding clarity in terms of underpinning explanations, applicability beyond typical development, and broader implications for socio-emotional development.

Overall, the current findings add clarity to the proposed cross-condition model of audio emotion recognition, highlighting the importance of developing ability within various components of the model. However, findings also suggested complexity, with possible between-condition variation in the ways in which certain model components operate and interact during emotion recognition. This calls for further developmental research, including involving children with socio-emotional difficulties, to elucidate the aspects within which music and voice converge and diverge. Further, research replicating music-specific associations between emotion recognition and externalising difficulties and exploring possible condition-general associations with a greater range of socio-emotional dimensions would reveal the broader developmental implications of these converging and diverging patterns. Crucially, this would identify suitable outcomes and mechanistic targets for music-based interventions, either via condition-general or music-specific aspects of the proposed audio emotion recognition model.

References

Achenbach, T. M., & Edelbrock, C. S. (1978). The classification of child psychopathology: a review and analysis of empirical efforts. *Psychological bulletin*, *85*(6), 1275.

Achenbach, T. M., Ivanova, M. Y., Rescorla, L. A., Turner, L. V., & Althoff, R. R. (2016). Internalizing/externalizing problems: Review and recommendations for clinical and research applications. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(8), 647-656.

Adachi, M., & Trehub, S. E. (1998). Children's expression of emotion in song. *Psychology of Music*, *26*(2), 133-153.

Airdrie, J. N., Langley, K., Thapar, A., & van Goozen, S. H. (2018). Facial emotion recognition and eye gaze in attention-deficit/hyperactivity disorder with and without comorbid conduct disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *57*(8), 561-570.

Allgood, R., & Heaton, P. (2015). Developmental change and cross-domain links in vocal and musical emotion recognition performance in childhood. *British Journal of Developmental Psychology*, *33*(3), 398-403.

Alviar, C., Sahoo, M., Edwards, L. A., Jones, W., Klin, A., & Lense, M. (2023). Infant-directed song potentiates infants' selective attention to adults' mouths over the first year of life. *Developmental science*, *26*(5), e13359.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388-407.

Argstatter, H. (2016). Perception of basic emotions in music: Culture-specific or multicultural? *Psychology of Music*, *44*(4), 674-690.

Astle, D. E., Holmes, J., Kievit, R., & Gathercole, S. E. (2022). Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders. *Journal of Child Psychology and Psychiatry*, *63*(4), 397-417.

Athanasopoulos, G., Eerola, T., Lahdelma, I., & Kaliakatsos-Papakostas, M. (2021). Harmonic organisation conveys both universal and culture-specific cues for emotional expression in music. *PloS one*, *16*(1), e0244964.

Balkwill, L.-L., & Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception*, *17*(1), 43-64.

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

Bänziger, T., Hosoya, G., & Scherer, K. R. (2015). Path models of vocal emotion communication. *PloS one*, *10*(9), e0136675.

Baron-Cohen, S., Golan, O., Wheelwright, S., Granader, Y., & Hill, J. (2010). Emotion word comprehension from 4 to 16 years old: A developmental survey. *Frontiers in evolutionary neuroscience*, *2*, 109.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, *1*(1), 28-58.

Barrett, L. F. (2013). Psychological construction: The Darwinian approach to the science of emotion. *Emotion Review*, *5*(4), 379-389.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1-23.

Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in experimental social psychology*, *41*, 167-218.

Bartsch, K. (1995). Children talk about the mind. *Oxford University Press google schola*, *2*, 228-247.

Beck, L., Kumschick, I. R., Eid, M., & Klann-Delius, G. (2012). Relationship between language competence and emotional competence in middle childhood. *Emotion*, *12*(3), 503.

Bedoya, D., Arias, P., Rachman, L., Liuni, M., Canonne, C., Goupil, L., & Aucouturier, J.-J. (2021). Even violins can cry: specifically vocal emotional behaviours also drive the perception of emotions in non-vocal music. *Philosophical Transactions of the Royal Society B*, *376*(1840), 20200396.

Behavioral, D. o., Sciences, S., Research, C. o. C. D., Policy, P., & Children, P. t. R. t. S. o. B. R. o. S.-A. (1984). Development during middle childhood: The years from six to twelve.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2002). Voice-selective areas in human auditory cortex.

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child development*, *89*(6), 1996-2009.

Bestelmeyer, P. E., Maurage, P., Rouger, J., Latinus, M., & Belin, P. (2014). Adaptation to vocal expressions reveals multistep perception of auditory emotion. *Journal of Neuroscience*, *34*(24), 8098-8105.

Blasco-Magraner, J. S., Bernabe-Valero, G., Marín-Liébana, P., & Moret-Tatay, C. (2021). Effects of the Educational Use of Music on 3-to 12-Year-Old Children's Emotional Development: A Systematic Review. *International journal of environmental research and public health*, *18*(7), 3668.

Bliss-Moreau, E., Williams, L. A., & Santistevan, A. C. (2020). The immutability of valence and arousal in the foundation of emotion. *Emotion*, *20*(6), 993.

Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences*, *98*(20), 11818-11823.

Boone, R. T., & Cunningham, J. G. (1998). Children's decoding of emotion in expressive body movement: the development of cue attunement. *Developmental Psychology*, *34*(5), 1007.

Bor, W., McGee, T. R., & Fagan, A. A. (2004). Early risk factors for adolescent antisocial behaviour: An Australian longitudinal study. *Australian & New Zealand Journal of Psychiatry*, *38*(5), 365-372.

Bosacki, S. L., & Moore, C. (2004). Preschoolers' understanding of simple and complex emotions: Links with gender and language. *Sex roles*, *50*, 659-675.

Boucher, H., Gaudette-Leblanc, A., Raymond, J., & Peters, V. (2021). Musical learning as a contributing factor in the development of socio-emotional competence in children aged 4 and 5: an Exploratory study in a naturalistic context. *Early Child Development and Care*, *191*(12), 1922-1938.

Boucher, J. (2000). Time parsing, normal language acquisition, and language-related developmental disorders. In *New directions in language development and disorders* (pp. 13-23). Springer.

Brooks, J. A., & Freeman, J. B. (2018). Conceptual knowledge predicts the representational structure of facial emotion perception. *Nature human behaviour*, *2*(8), 581-591.

Brown, S. (2017). A joint prosodic origin of language and music. *Frontiers in psychology*, *8*, 1894.

Bruce Morton, J., & Trehub, S. E. (2007). Children's judgements of emotion in song. *Psychology of Music*, *35*(4), 629-639.

Bryant, G., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture*, *8*(1-2), 135-148.

Brysbaert, M. (2024). Designing and evaluating tasks to measure individual differences in experimental psychology: a tutorial. *Cognitive Research: Principles and Implications*, *9*(1), 11.

Buck, R. (2014). *Emotion: A biosocial synthesis*. Cambridge University Press.

Buss, D. M. (2005). *The handbook of evolutionary psychology*. Wiley Online Library.

Cadesky, E. B., Mota, V. L., & Schachar, R. J. (2000). Beyond words: how do children with ADHD and/or conduct problems process nonverbal information about affect? *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(9), 1160-1167.

Cameron, C. D., Lindquist, K. A., & Gray, K. (2015). A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review*, *19*(4), 371-394.

Campos, J. J., Campos, R. G., & Barrett, K. C. (1989). Emergent themes in the study of emotional development and emotion regulation. *Developmental Psychology*, *25*(3), 394.

Castro, V. L., Cooke, A. N., Halberstadt, A. G., & Garrett-Peters, P. (2018). Bidirectional linkages between emotion recognition and problem behaviors in elementary school children. *Journal of nonverbal behavior*, *42*, 155-178.

Celeghin, A., Diano, M., Bagnis, A., Viola, M., & Tamietto, M. (2017). Basic emotions in human neuroscience: neuroimaging and beyond. *Frontiers in psychology*, *8*, 1432.

Cespedes-Guevara, J., & Eerola, T. (2018). Music communicates affects, not basic emotions–a constructionist account of attribution of emotional meanings to music. *Frontiers in psychology*, *9*, 215.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., De Rosario, H., & De Rosario, M. H. (2018). Package 'pwr'. *R package version*, *1*(2).

Chronaki, G., Garner, M., Hadwin, J. A., Thompson, M. J., Chin, C. Y., & Sonuga-Barke, E. J. (2015). Emotion-recognition abilities and behavior problem dimensions in preschoolers:

evidence for a specific role for childhood hyperactivity. *Child Neuropsychology*, *21*(1), 25-40.

Chronaki, G., Hadwin, J. A., Garner, M., Maurage, P., & Sonuga-Barke, E. J. (2015). The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. *British Journal of Developmental Psychology*, *33*(2), 218-236.

Chronaki, G., Wigelsworth, M., Pell, M. D., & Kotz, S. A. (2018). The development of cross-cultural recognition of vocal emotion during childhood and adolescence. *Scientific reports*, *8*(1), 1-17.

Cirelli, L. K., Jurewicz, Z. B., & Trehub, S. E. (2020). Effects of maternal singing style on mother–infant arousal and behavior. *Journal of cognitive neuroscience*, *32*(7), 1213-1220.

Clark, C. N., Downey, L. E., & Warren, J. D. (2015). Brain disorders and the biological role of music. *Social Cognitive and Affective Neuroscience*, *10*(3), 444-452.

Cole, P. M. (2016). Emotion and the development of psychopathology. *Developmental psychopathology*, 1-60.

Cooper, S., Hobson, C. W., & van Goozen, S. H. (2020). Facial emotion recognition in children with externalising behaviours: A systematic review. *Clinical child psychology and psychiatry*, *25*(4), 1068-1085.

Corbeil, M., Trehub, S. E., & Peretz, I. (2016). Singing delays the onset of infant distress. *Infancy*, *21*(3), 373-391.

Correia, A. I., Castro, S. L., MacGregor, C., Müllensiefen, D., Schellenberg, E. G., & Lima, C. F. (2020). Enhanced recognition of vocal emotions in individuals with naturally good musical abilities. *Emotion*.

Coutinho, E., & Dibben, N. (2013). Psychoacoustic cues to emotion in speech prosody and music. *Cognition & Emotion*, *27*(4), 658-684.

Coutinho, E., Scherer, K. R., & Dibben, N. (2014). Singing and emotion. *The Oxford handbook of singing*, 297-314.

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, *3*(4), 369-382.

Creighton, A. L., Atherton, M., Kitamura, C., & Trondalen, G. (2013). Singing play songs and lullabies: Investigating the subjective contributions to maternal attachment constructs. *Australian Journal of Music Therapy*, *24*, 17-47.

Cross, I. (2014). Music and communication in music psychology. *Psychology of Music*, *42*(6), 809-819.

Dalla Bella, S., Peretz, I., Rousseau, L., & Gosselin, N. (2001). A developmental study of the affective value of tempo and mode in music. *Cognition*, *80*(3), B1-B10.

Davis-Kean, P. E., & Ellis, A. (2019). An overview of issues in infant and developmental research for the creation of robust and replicable science. *Infant behavior and development*, *57*, 101339.

Dawel, A., O'Kearney, R., McKone, E., & Palermo, R. (2012). Not just fear and sadness: Meta-analytic evidence of pervasive emotion recognition deficits for facial and vocal expressions in psychopathy. *Neuroscience & Biobehavioral Reviews*, *36*(10), 2288-2304.

Day, R., & Thompson, W. F. (2025). How Does Music Elicit Emotions? In *Emotion Theory: The Routledge Comprehensive Guide* (pp. 407-422). Routledge.

de Villiers, J. G., & de Villiers, P. A. (2003). Language for thought: Coming to understand false beliefs.

Dede, B., Delk, L., & White, B. A. (2021). Relationships between facial emotion recognition, internalizing symptoms, and social problems in young children. *Personality and Individual Differences*, *171*, 110448.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *science*, *284*(5416), 970-974.

Demetriou, C. A., Colins, O. F., Andershed, H., & Fanti, K. A. (2023). Assessing psychopathic traits early in development: Testing potential associations with social, behavioral, and affective factors. *Journal of Psychopathology and Behavioral Assessment*, *45*(3), 767-780.

Denham, S. (1998). *Emotional development in young children*. Guilford Press.

Dolgin, K. G., & Adelson, E. H. (1990). Age changes in the ability to interpret affect in sung and instrumentally-presented melodies. *Psychology of Music*, *18*(1), 87-98.

Donhauser, P. W., & Klein, D. (2022). Audio-Tokens: A toolbox for rating, sorting and comparing audio samples in the browser. *Behavior Research Methods*, 1-8.

Dunn, L. M., & Dunn, D. M. (2009). *The British picture vocabulary scale*. GL Assessment Limited.

Durgungoz, F. C., & St Clair, M. C. (2024). An interactive technology-based emotion recognition intervention for children with developmental language disorder: A longitudinal mixed-method study. *European Journal of Special Needs Education*, 1-17.

Easter, J., McClure, E. B., Monk, C. S., Dhanani, M., Hodgdon, H., Leibenluft, E., Charney, D. S., Pine, D. S., & Ernst, M. (2005). Emotion recognition deficits in pediatric anxiety disorders: Implications for amygdala research. *Journal of Child & Adolescent Psychopharmacology*, *15*(4), 563-570.

Eaton, N. R., Rodriguez-Seijas, C., Carragher, N., & Krueger, R. F. (2015). Transdiagnostic factors of psychopathology and substance use disorders: a review. *Social psychiatry and psychiatric epidemiology*, *50*, 171-182.

Ecklund-Flores, L., & Turkewitz, G. (1996). Asymmetric headturning to speech and nonspeech in human newborns. *Developmental psychobiology*, *29*(3), 205-217.

Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: contribution, linearity, and additivity of primary musical cues. *Frontiers in psychology*, *4*, 487.

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*(1), 18-49.

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review*, *3*(4), 364-370.

Elfenbein, H. A., Laukka, P., Althoff, J., Chui, W., Iraki, F. K., Rockstuhl, T., & Thingujam, N. S. (2021). What Do We Hear in the Voice? An Open-Ended Judgment Study of Emotional Speech Prosody. *Personality and Social Psychology Bulletin*, 01461672211029786.

Ellsworth, P. C. (2013). Appraisal theory: Old and new questions. *Emotion Review*, *5*(2), 125-131.

Eme, R. (2017). Developmental psychopathology: A primer for clinical pediatrics. *World journal of psychiatry*, *7*(3), 159.

Escoffier, N., Zhong, J., Schirmer, A., & Qiu, A. (2013). Emotional expressions in voice and music: same code, same effect? *Human brain mapping*, *34*(8), 1796-1810.

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

Ezard, G., Slack, J., Pearce, M. J., & Hodgson, T. L. (2022). Applying the British picture vocabulary scale to estimate premorbid cognitive ability in adults. *Applied Neuropsychology: Adult*, *29*(5), 1049-1059.

Fancourt, D., & Perkins, R. (2018). Maternal engagement with music up to nine months post-birth: Findings from a cross-sectional study in England. *Psychology of Music*, *46*(2), 238-251.

Fedorenko, E., Patel, A., Casasanto, D., Winawer, J., & Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory & cognition*, *37*, 1-9.

Flom, R., & Bahrick, L. E. (2007). The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology*, *43*(1), 238.

Franco, F., Chew, M., & Swaine, J. S. (2017). Preschoolers' attribution of affect to music: A comparison between vocal and instrumental performance. *Psychology of Music*, *45*(1), 131-149.

Franco, F., Suttora, C., Spinelli, M., Kozar, I., & Fasolo, M. (2022). Singing to infants matters: Early singing interactions affect musical preferences and facilitate vocabulary building. *Journal of Child Language*, *49*(3), 552-577.

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A. D., & Koelsch, S. (2009). Universal recognition of three basic emotions in music. *Current Biology*, *19*(7), 573-576.

Fritz, T. H., Schütte, F., Steixner, A., Contier, O., Obrig, H., & Villringer, A. (2019). Musical meaning modulates word acquisition. *Brain and language*, *190*, 10-15.

Frühholz, S., Trost, W., & Kotz, S. A. (2016). The sound of emotions—Towards a unifying neural network perspective of affective sound processing. *Neuroscience & Biobehavioral Reviews*, *68*, 96-110.

Fuentes-Sanchez, N., Pastor, M. C., Eerola, T., & Pastor, R. (2023). Individual differences in music reward sensitivity influence the perception of emotions represented by music. *Musicae Scientiae*, *27*(2), 313-331.

Gabrielsson, A., & Lindström, E. (2010). The role of structure in the musical expression of emotions. *Handbook of music and emotion: Theory, research, applications*, *367400*, 367-344.

Gao, X., & Maurer, D. (2010). A happy story: Developmental changes in children's sensitivity to facial expressions of varying intensities. *Journal of experimental child psychology*, *107*(2), 67-86.

Gatzke-Kopp, L. M., Beauchaine, T. P., Shannon, K. E., Chipman, J., Fleming, A. P., Crowell, S. E., Liang, O., Johnson, L. C., & Aylward, E. (2009). Neurological correlates of reward

responding in adolescents with and without externalizing behavior disorders. *Journal of Abnormal Psychology*, *118*(1), 203.

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. S. (1993). Comparing internal and external standards in voice quality judgments. *Journal of speech, language, and hearing research*, *36*(1), 14-20.

Giordano, B. L., Whiting, C., Kriegeskorte, N., Kotz, S. A., Gross, J., & Belin, P. (2021). The representational dynamics of perceived voice emotions evolve from categories to dimensions. *Nature human behaviour*, 1-11.

Gold, C., Voracek, M., & Wigram, T. (2004). Effects of music therapy for children and adolescents with psychopathology: a meta-analysis. *Journal of Child Psychology and Psychiatry*, *45*(6), 1054-1063.

Gomez, P., & Danuser, B. (2007). Relationships between musical structure and psychophysiological measures of emotion. *Emotion*, *7*(2), 377.

Gómez-Cañón, J. S., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y.-H., & Gómez, E. (2021). Music Emotion Recognition: Toward new, robust standards in personalized and context-sensitive applications. *IEEE Signal Processing Magazine*, *38*(6), 106-114.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581-586.

Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, *128*(3), 1322-1336.

Grandjean, D., Bänziger, T., & Scherer, K. R. (2006). Intonation as an interface between language and affect. *Progress in brain research*, *156*, 235-247.

Graziano, P. A., & Garcia, A. (2016). Attention-deficit hyperactivity disorder and children's emotion dysregulation: A meta-analysis. *Clinical psychology review*, *46*, 106-123.

Grazzani, I., & Ornaghi, V. (2012). How do use and comprehension of mental-state language relate to theory of mind in middle childhood? *Cognitive development*, *27*(2), 99-111.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493-498.

Griffiths, S., Goh, S. K. Y., & Norbury, C. F. (2020). Early language competence, but not general cognitive ability, predicts children's recognition of emotion from facial and vocal cues. *PeerJ*, *8*, e9118.

Grosbras, M.-H., Ross, P. D., & Belin, P. (2018). Categorical emotion recognition from voice improves during childhood and adolescence. *Scientific reports*, *8*(1), 1-11.

Grosse, G., & Streubel, B. (2024). Emotion-specific vocabulary and its relation to emotion understanding in children and adolescents. *Cognition and emotion*, 1-10.

Grossi, G., Strappini, F., Iuliano, E., Passiatore, Y., Mancini, F., Levantini, V., Masi, G., Milone, A., Santaguida, E., & Salekin, R. T. (2023). Psychopathic traits, externalizing problems, and prosocial behavior: The role of social dominance orientation. *Journal of Clinical Medicine*, *12*(10), 3521.

Grossmann, T., Oberecker, R., Koch, S. P., & Friederici, A. D. (2010). The developmental origins of voice processing in the human brain. *Neuron*, *65*(6), 852-858.

Hailstone, J. C., Omar, R., Henley, S. M., Frost, C., Kenward, M. G., & Warren, J. D. (2009). It's not what you play, it's how you play it: Timbre affects perception of emotion in music. *Quarterly Journal of Experimental Psychology*, *62*(11), 2141-2155.

Hannant, P. (2018). Receptive language is associated with visual perception in typically developing children and sensorimotor skills in autism spectrum conditions. *Human movement science*, *58*, 297-306.

Harmon-Jones, E., Harmon-Jones, C., & Summerell, E. (2017). On the importance of both dimensional and discrete models of emotion. *Behavioral Sciences*, *7*(4), 66.

Harris, P. L., de Rosnay, M., & Pons, F. (2005). Language and children's understanding of mental states. *Current Directions in Psychological Science*, *14*(2), 69-73.

Haslam, N., McGrath, M. J., Viechtbauer, W., & Kuppens, P. (2020). Dimensions over categories: A meta-analysis of taxometric research. *Psychological medicine*, *50*(9), 1418-1432.

Herrando, C., & Constantinides, E. (2021). Emotional Contagion: A Brief Overview and Future Directions. *Frontiers in psychology*, 2881.

Holz, N., Larrouy-Maestri, P., & Poeppel, D. (2021). The paradoxical role of emotional intensity in the perception of vocal affect. *Scientific reports*, *11*(1), 1-10.

Hunter, P. G., Schellenberg, E. G., & Stalinski, S. M. (2011). Liking and identifying emotionally expressive music: Age and gender differences. *Journal of experimental child psychology*, *110*(1), 80-93.

Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, *23*(4), 319-330.

Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, *23*(4), 319-330.

Insel, T. R. (2014). The NIMH research domain criteria (RDoC) project: precision medicine for psychiatry. *American Journal of Psychiatry*, *171*(4), 395-397.

Izard, C. E. (2009). Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Annual review of psychology*, *60*, 1-25.

Jansen, N., Harding, E. E., Loerts, H., Başkent, D., & Lowie, W. (2023). The relation between musical abilities and speech prosody perception: A meta-analysis. *Journal of Phonetics*, *101*, 101278.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Jentschke, S. (2014). The relationship between music and language.

Jones, C. R., Pickles, A., Falcaro, M., Marsden, A. J., Happé, F., Scott, S. K., Sauter, D., Tregay, J., Phillips, R. J., & Baird, G. (2011). A multimodal approach to emotion recognition ability in autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, *52*(3), 275-285.

Juslin, P. N. (1997). Emotional communication in music performance: A functionalist perspective and some data. *Music Perception*, *14*(4), 383-418.

Juslin, P. N. (2013). What does music express? Basic emotions and beyond. *Frontiers in psychology*, *4*, 596.

Juslin, P. N. (2018). *Musical emotions explained: Unlocking the secrets of musical affect*. Oxford University Press, USA.

Juslin, P. N., Harmat, L., & Eerola, T. (2014). What makes music emotionally significant? Exploring the underlying mechanisms. *Psychology of Music*, *42*(4), 599-623.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, *129*(5), 770.

Juslin, P. N., & Lindström, E. (2010). Musical expression of emotions: Modelling listeners' judgements of composed and performed features. *Music Analysis*, *29*(1-3), 334-364.

Juslin, P. N., & Sloboda, J. (2011). *Handbook of music and emotion: Theory, research, applications*. Oxford University Press.

Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic bulletin & review*, *27*(2), 237-265.

Kasparek, S. W., Jenness, J. L., & McLaughlin, K. A. (2020). Reward processing modulates the association between trauma exposure and externalizing psychopathology. *Clinical Psychological Science*, *8*(6), 989-1006.

Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior*, 1-28.

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature neuroscience*, *23*(7), 788-799.

Kim, H.-S., & Kim, H.-S. (2018). Effect of a musical instrument performance program on emotional intelligence, anxiety, and aggression in Korean elementary school children. *Psychology of Music*, *46*(3), 440-453.

Koelsch, S. (2014). Brain correlates of music-evoked emotions. *Nature Reviews Neuroscience*, *15*(3), 170-180.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. D. (2004). Music, language and meaning: brain signatures of semantic processing. *Nature neuroscience*, *7*(3), 302-307.

Korhonen, M., Luoma, I., Salmelin, R., Siirtola, A., & Puura, K. (2018). The trajectories of internalizing and externalizing problems from early childhood to adolescence and young adult outcome.

Kragness, H. E., Eitel, M. J., Baksh, A. M., & Trainor, L. J. (2021). Evidence for early arousal-based differentiation of emotions in children's musical production. *Developmental science*, *24*(1), e12982.

Kratus, J. (1993). A developmental study of children's interpretation of emotion in music. *Psychology of Music*, *21*(1), 3-19.

Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., Simms, L. J., Widiger, T. A., Achenbach, T. M., & Bach, B. (2018). Progress in achieving quantitative classification of psychopathology. *World Psychiatry*, *17*(3), 282-293.

Lahey, B. B., Moore, T. M., Kaczkurkin, A. N., & Zald, D. H. (2021). Hierarchical models of psychopathology: Empirical support, implications, and remaining issues. *World Psychiatry*, *20*(1), 57-63.

Lahey, B. B., Tiemeier, H., & Krueger, R. F. (2022). Seven reasons why binary diagnostic categories should be replaced with empirically sounder and less stigmatizing dimensions. *JCPP advances*, *2*(4), e12108.

Landerl, K., & Moll, K. (2010). Comorbidity of learning disorders: prevalence and familial transmission. *Journal of Child Psychology and Psychiatry*, *51*(3), 287-294.

Laukka, P., Eerola, T., Thingujam, N. S., Yamasaki, T., & Beller, G. (2013). Universal and culture-specific factors in the recognition and performance of musical affect expressions. *Emotion*, *13*(3), 434.

Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Iraki, F. K. e., Rockstuhl, T., & Thingujam, N. S. (2013). Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in psychology*, *4*, 353.

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, *19*(5), 633-653.

Laukka, P., & Juslin, P. N. (2007). Similar patterns of age-related differences in emotion recognition from speech and music. *Motivation and emotion*, *31*, 182-191.

Lausen, A., & Hammerschmidt, K. (2020). Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications*, *7*(1), 1-17.

Lemerise, E. A., & Arsenio, W. F. (2000). An integrated model of emotion processes and cognition in social information processing. *Child development*, *71*(1), 107-118.

Lense, M. D., Shultz, S., Astésano, C., & Jones, W. (2022). Music of infant-directed singing entrains infants' social visual behavior. *Proceedings of the National Academy of Sciences*, *119*(45), e2116967119.

Lima, C. F., Brancatisano, O., Fancourt, A., Müllensiefen, D., Scott, S. K., Warren, J. D., & Stewart, L. (2016). Impaired socio-emotional processing in a developmental music disorder. *Scientific reports*, *6*(1), 1-13.

Lima, C. F., & Castro, S. L. (2011). Speaking to the trained ear: musical expertise enhances the recognition of emotions in speech prosody. *Emotion*, *11*(5), 1021.

Lindquist, K. A. (2017). The role of language in emotion: existing evidence and future directions. *Current opinion in psychology*, *17*, 135-139.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences*, *35*(3), 121-143.

Lindström, E. (2006). Impact of melodic organization on perceived structure and emotional expression in music. *Musicae Scientiae*, *10*(1), 85-117.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, *13*(5), e0196391.

LoBue, V., Pérez-Edgar, K., & Buss, K. A. (2019). *Handbook of emotional development*. Springer.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of research in Personality*, *48*, 61-83.

Markman, A. B., & Rein, J. R. (2013). The nature of mental concepts. *The Oxford Handbook of Cognitive Psychology*, 321-329.

Martins, M., Pinheiro, A. P., & Lima, C. F. (2020). Does Music Training Improve Emotion Recognition Abilities? A Critical Review. *Emotion Review*, 17540739211022035.

Masataka, N. (2009). The origins of language and the evolution of music: A comparative perspective. *Physics of life reviews*, *6*(1), 11-22.

McDermott, J., & Hauser, M. (2005). The origins of music: Innateness, uniqueness, and evolution. *Music Perception*, *23*(1), 29-59.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, *17*(3), 437.

Micallef Grimaud, A., & Eerola, T. (2022). An interactive approach to emotional expression through musical cues. *Music & Science*, *5*, 20592043211061745.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551-1562.

Mohn, C., Argstatter, H., & Wilker, F.-W. (2011). Perception of six basic emotions in music. *Psychology of Music*, *39*(4), 503-517.

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, *5*(2), 119-124.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6-18.

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'bayesfactor'. In.

Morningstar, M., Nelson, E. E., & Dirks, M. A. (2018). Maturation of vocal emotion recognition: Insights from the developmental and neuroimaging literature. *Neuroscience & Biobehavioral Reviews*, *90*, 221-230.

Mosing, M. A., Peretz, I., & Ullén, F. (2017). Genetic influences on music expertise. *The science of expertise*, 272-282.

Mote, J. (2011). The effects of tempo and familiarity on children's affective interpretation of music. *Emotion*, *11*(3), 618.

Mualem, O., & Lavidor, M. (2015). Music education intervention improves vocal emotion recognition. *International Journal of Music Education*, *33*(4), 413-425.

Mullin, B. C., & Hinshaw, S. P. (2007). Emotion Regulation and Externalizing Disorders in Children and Adolescents.

Muratori, P., Buonanno, C., Gallani, A., Grossi, G., Levantini, V., Milone, A., Pisano, S., Salekin, R. T., Sesso, G., & Masi, G. (2021). Validation of the Proposed Specifiers for Conduct Disorder (PSCD) scale in a sample of Italian students. *Children*, *8*(11), 1020.

Nakata, T., & Trehub, S. E. (2004). Infants' responsiveness to maternal speech and singing. *Infant behavior and development*, *27*(4), 455-464.

Nakata, T., & Trehub, S. E. (2011). Expressive timing and dynamics in infant-directed and non-infant-directed singing. *Psychomusicology: Music, Mind and Brain*, *21*(1-2), 45.

Nawrot, E. S. (2003). The perception of emotional expression in music: Evidence from infants, children and adults. *Psychology of Music*, *31*(1), 75-92.

Nelson, N. L., & Russell, J. A. (2011). Preschoolers' use of dynamic facial, bodily, and vocal cues to emotion. *Journal of experimental child psychology*, *110*(1), 52-61.

Nencheva, M. L., Nook, E., Thornton, M. A., Lew-Williams, C., & Tamir, D. (2023). The co-emergence of emotion vocabulary and organized emotion dynamics in childhood.

Neves, L., Martins, M., Correia, A. I., Castro, S. L., & Lima, C. (2021). Associations Between Vocal Emotion Recognition and Socio-emotional Adjustment in Children. *bioRxiv*.

Nook, E. C., Lindquist, K. A., & Zaki, J. (2015). A new look at emotion perception: Concepts speed and shape facial emotion recognition. *Emotion*, *15*(5), 569.

Nook, E. C., Sasse, S. F., Lambert, H. K., McLaughlin, K. A., & Somerville, L. H. (2017). Increasing verbal knowledge mediates development of multidimensional emotion representations. *Nature human behaviour*, *1*(12), 881-889.

Nook, E. C., Stavish, C. M., Sasse, S. F., Lambert, H. K., Mair, P., McLaughlin, K. A., & Somerville, L. H. (2020). Charting the development of emotion comprehension and abstraction from childhood to adulthood using observer-rated and linguistic measures. *Emotion*, *20*(5), 773.

Nordström, H., & Laukka, P. (2019). The time course of emotion recognition in speech and music. *The Journal of the Acoustical Society of America*, *145*(5), 3058-3074.

Nowicki Jr, S., & Mitchell, J. (1998). Accuracy in identifying affect in child and adult faces and voices and social competence in preschool children. *Genetic, Social, and General Psychology Monographs*, *124*(1), 39-60.

Nowicki, S., Bliwise, N., & Joinson, C. (2019). The association of children's locus of control orientation and emotion recognition abilities at 8 years of age and teachers' ratings of their personal and social difficulties at 10 years. *Journal of nonverbal behavior*, *43*, 381-396.

Nussbaum, C., & Schweinberger, S. R. (2021). Links between Musicality and Vocal Emotion Perception.

Oatley, K. (1992). *Best laid schemes: The psychology of the emotions*. Cambridge University Press.

Oberauer, K. (2019). Working memory capacity limits memory for bindings. *Journal of Cognition*, *2*(1).

Oberauer, K. (2023). Measurement models for visual working memory—A factorial model comparison. *Psychological Review*, *130*(3), 841.

Ornaghi, V., & Grazzani, I. (2013). The relationship between emotional-state language and emotion understanding: A study with school-age children. *Cognition & Emotion*, *27*(2), 356-366.

Pandolfi, V., Magyar, C. I., & Dill, C. A. (2012). An initial psychometric evaluation of the CBCL 6–18 in a sample of youth with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *6*(1), 96-108.

Papachristou, E., & Flouri, E. (2020). Distinct developmental trajectories of internalising and externalising symptoms in childhood: Links with mental health and risky behaviours in early adolescence. *Journal of Affective Disorders*, *276*, 1052-1060.

Papadimitriou, A., Smyth, C., Politimou, N., Franco, F., & Stewart, L. (2021). The impact of the home musical environment on infants' language development. *Infant behavior and development*, *65*, 101651.

Paquette, S., Ahmed, G., Goffi-Gomez, M., Hoshino, A., Peretz, I., & Lehmann, A. (2018). Musical and vocal emotion perception for cochlear implants users. *Hearing research*, *370*, 272-282.

Paquette, S., Peretz, I., & Belin, P. (2013). The "Musical Emotional Bursts": a validated set of musical affect bursts to investigate auditory affective processing. *Frontiers in psychology*, *4*, 509.

Patel, A. D. (2010). *Music, language, and the brain*. Oxford university press.

Peretz, I., Vuvan, D., Lagrois, M.-É., & Armony, J. L. (2015). Neural overlap in processing music and speech. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1664), 20140090.

Persico, G., Antolini, L., Vergani, P., Costantini, W., Nardi, M. T., & Bellotti, L. (2017). Maternal singing of lullabies during pregnancy and after birth: Effects on mother–infant bonding and on newborns' behaviour. Concurrent cohort study. *Women and Birth*, *30*(4), e214-e220.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *science*, *306*(5695), 499-503.

Pinker, S. (1997). How the mind works. *WW Norton&Company*.

Pino, M. C., Giancola, M., & D'Amico, S. (2023). The association between music and language in children: A state-of-the-art review. *Children*, *10*(5), 801.

Plate, R. C., Jones, C., Steinberg, J. S., Daley, G., Zhao, S., & Waller, R. (2022). Alignment of Knowing Versus Feeling the Emotion in Music During Middle-Childhood. Proceedings of the Annual Meeting of the Cognitive Science Society,

Politimou, N., Stewart, L., Müllensiefen, D., & Franco, F. (2018). Music@ Home: A novel instrument to assess the home musical environment in the early years. *PloS one*, *13*(4), e0193819.

Pons, F., Lawson, J., Harris, P. L., & De Rosnay, M. (2003). Individual differences in children's emotion understanding: Effects of age and language. *Scandinavian journal of psychology*, *44*(4), 347-353.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, *17*(3), 715-734.

Proverbio, A. M., Camporeale, E., & Brusa, A. (2020). Multimodal recognition of emotions in music and facial expressions. *Frontiers in human neuroscience*, *14*.

Proverbio, A. M., De Benedetto, F., & Guazzone, M. (2020). Shared neural mechanisms for processing emotions in music and vocalizations. *European Journal of Neuroscience*, *51*(9), 1987-2007.

Proverbio, A. M., & Piotti, E. (2022). Common neural bases for processing speech prosody and music: An integrated model. *Psychology of Music*, *50*(5), 1408-1423.

Quintana, D. S., & Williams, D. R. (2018). Bayesian alternatives for common null-hypothesis significance tests in psychiatry: a non-technical guide using JASP. *BMC psychiatry*, *18*(1), 1-8.

Reybrouck, M., & Podlipniak, P. (2019). Preconceptual spectral and temporal cues as a source of meaning in speech and music. *Brain sciences*, *9*(3), 53.

Rose, D. C., Heaton, P., & Bartoli, A. J. (2015). Changes in the wellbeing of children starting to learn to play musical instruments. *Assessment & Development Matters*, *7*(1), 26-30.

Rosen, P. J., Leaberry, K. D., Slaughter, K., Fogleman, N. D., Walerius, D. M., Loren, R. E., & Epstein, J. N. (2019). Managing Frustration for Children (MFC) group intervention for

ADHD: An open trial of a novel group intervention for deficient emotion regulation. *Cognitive and Behavioral practice*, *26*(3), 522-534.

Rosenqvist, J., Lahti-Nuuttila, P., Laasonen, M., & Korkman, M. (2014). Preschoolers' recognition of emotional expressions: Relationships with other neurocognitive capacities. *Child Neuropsychology*, *20*(3), 281-302.

Ross, B. H., & Spalding, T. L. (1994). Concepts and categories. In *Thinking and problem solving* (pp. 119-148). Elsevier.

Ross, P., Atkins, B., Allison, L., Simpson, H., Duffell, C., Williams, M., & Ermolina, O. (2021). Children cannot ignore what they hear: Incongruent emotional information leads to an auditory dominance in children. *Journal of experimental child psychology*, *204*, 105068.

Ross, P., Williams, E., Herbert, G., Manning, L., & Lee, B. (2023). Turn that music down! Affective musical bursts cause an auditory dominance in children recognizing bodily emotions. *Journal of experimental child psychology*, *230*, 105632.

Ruba, A. L., & Pollak, S. D. (2020). The development of emotion reasoning in infancy and early childhood. *Annual Review of Developmental Psychology*, *2*, 503-531.

Ruba, A. L., & Repacholi, B. M. (2020). Beyond language in infant emotion concept development. *Emotion Review*, *12*(4), 255-258.

Ruba, A. L., Wilbourn, M. P., Ulrich, D. M., & Harris, L. T. (2018). Constructing emotion categorization: Insights from developmental psychology applied to a young adult sample. *Emotion*, *18*(7), 1043.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145.

Russell, J. A., & Widen, S. C. (2002). A label superiority effect in children's categorization of facial expressions. *Social development*, *11*(1), 30-52.

Saarikallio, S. (2009). Emotional self-regulation through music in 3-8-year-old children.

Saarikallio, S. (2011). Music as emotional self-regulation throughout adulthood. Psychology of music, 39(3), 307-327.

Saarikallio, S., Vuoskoski, J., & Luck, G. (2014). Adolescents' expression and perception of emotion in music reflects their broader abilities of emotional communication. *Psychology of Well-Being*, *4*(1), 21.

Saarni, C. (1999). *The development of emotional competence*. Guilford press.

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., Laznik, M.-C., & Cohen, D. (2013). Motherese in interaction: at the cross-road of emotion and cognition?(A systematic review). *PloS one*, *8*(10), e78103.

Sanfilippo, K. R. M., Stewart, L., & Glover, V. (2021). How music may support perinatal mental health: an overview. *Archives of women's mental health*, *24*(5), 831-839.

Satpute, A. B., & Lindquist, K. A. (2021). At the neural intersection between language and emotion. *Affective Science*, *2*(2), 207-220.

Satpute, A. B., Nook, E. C., Narayanan, S., Shu, J., Weber, J., & Ochsner, K. N. (2016). Emotions in "black and white" or shades of gray? How we think about emotion shapes our perception and neural representation of emotion. *Psychological science*, *27*(11), 1428-1442.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology*, *63*(11), 2251-2272.

Sauter, D. A., Panattoni, C., & Happé, F. (2013). Children's recognition of emotions from vocal cues. *British Journal of Developmental Psychology*, *31*(1), 97-113.

Schellenberg, E. G., Corrigall, K. A., Dys, S. P., & Malti, T. (2015). Group music training and children's prosocial skills. *PloS one*, *10*(10), e0141449.

Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of voice*, *9*(3), 235-248.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, *40*(1-2), 227-256.

Scherer, K. R. (2022). Theory convergence in emotion science is timely and realistic. *Cognition and emotion*, *36*(2), 154-170.

Scherer, K. R., Sundberg, J., Fantini, B., Trznadel, S., & Eyben, F. (2017). The expression of emotion in the singing voice: Acoustic patterns in vocal performance. *The Journal of the Acoustical Society of America*, *142*(4), 1805-1815.

Scherer, K. R., Sundberg, J., Tamarit, L., & Salomão, G. L. (2015). Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language*, *29*(1), 218-235.

Schirmer, A., & Adolphs, R. (2017). Emotion perception from face, voice, and touch: comparisons and convergence. *Trends in cognitive sciences*, *21*(3), 216-228.

Schirmer, A., Fox, P. M., & Grandjean, D. (2012). On the spatial organization of sound processing in the human temporal lobe: a meta-analysis. *Neuroimage*, *63*(1), 137-147.

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in cognitive sciences*, *10*(1), 24-30.

Schlegel, K., Palese, T., Mast, M. S., Rammsayer, T. H., Hall, J. A., & Murphy, N. A. (2020). A meta-analysis of the relationship between emotion recognition ability and intelligence. *Cognition and emotion*, *34*(2), 329-351.

Schön, D., Boyer, M., Moreno, S., Besson, M., Peretz, I., & Kolinsky, R. (2008). Songs as an aid for language acquisition. *Cognition*, *106*(2), 975-983.

Schubert, E., & McPherson, G. E. (2015). Underlying mechanisms and processes in the development of emotion perception in music. *The child as musician: A handbook of musical development*, 221-243.

Schultz, D., Izard, C. E., Ackerman, B. P., & Youngstrom, E. A. (2001). Emotion knowledge in economically disadvantaged children: Self-regulatory antecedents and relations to social difficulties and withdrawal. *Development and psychopathology*, *13*(1), 53-67.

Sells, R. C., Liversedge, S. P., & Chronaki, G. (2024). Vocal emotion recognition in attention-deficit hyperactivity disorder: a meta-analysis. *Cognition and emotion*, *38*(1), 23-43.

Shablack, H., & Lindquist, K. A. (2019). The role of language in emotional development. In *Handbook of emotional development* (pp. 451-478). Springer.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, *52*(6), 1061.

Shenfield, T., Trehub, S. E., & Nakata, T. (2003). Maternal singing modulates infant arousal. *Psychology of Music*, *31*(4), 365-375.

Simons, R. C. (1996). *Boo!: Culture, experience, and the startle reflex*. Oxford University Press, USA.

Siu, T.-S. C., & Cheung, H. (2017). Infants' sensitivity to emotion in music and emotion-action understanding. *PloS one*, *12*(2), e0171023.

Sivathasan, S., Dahary, H., Burack, J. A., & Quintin, E.-M. (2023). Basic emotion recognition of children on the autism spectrum is enhanced in music and typical for faces and voices. *PloS one*, *18*(1), e0279002.

Stachó, L., Saarikallio, S., Van Zijl, A., Huotilainen, M., & Toiviainen, P. (2013). Perception of emotional content in musical performances by 3–7-year-old children. *Musicae Scientiae*, *17*(4), 495-512.

Stewart, L., von Kriegstein, K., Warren, J. D., & Griffiths, T. D. (2006). Music and the brain: disorders of musical listening. *Brain*, *129*(10), 2533-2553.

Stewart, L., & Walsh, V. (2005). Infant learning: music and the baby brain. *Current Biology*, *15*(21), R882-R884.

Streubel, B., Gunzenhauser, C., Grosse, G., & Saalbach, H. (2020). Emotion-specific vocabulary and its contribution to emotion understanding in 4-to 9-year-old children. *Journal of experimental child psychology*, *193*, 104790.

Sturrock, A., & Freed, J. (2023). Preliminary data on the development of emotion vocabulary in typically developing children (5–13 years) using an experimental psycholinguistic measure. *Frontiers in psychology*, *13*, 982676.

Sutcliffe, R., Rendell, P. G., Henry, J. D., Bailey, P. E., & Ruffman, T. (2017). Music to my ears: Age-related decline in musical and facial emotion recognition. *Psychology and Aging*, *32*(8), 698.

Tan, Y. T., McPherson, G. E., Peretz, I., Berkovic, S. F., & Wilson, S. J. (2014). The genetic basis of music ability. *Frontiers in psychology*, *5*, 658.

Taylor, L. J., Maybery, M. T., Grayndler, L., & Whitehouse, A. J. (2015). Evidence for shared deficits in identifying emotions from faces and from voices in autism spectrum disorders and specific language impairment. *International journal of language & communication disorders*, *50*(4), 452-466.

Thiessen, E. D., & Saffran, J. R. (2009). How the melody facilitates the message and vice versa in infant learning and memory. *Annals of the New York Academy of Sciences*, *1169*(1), 225-233.

Thompson, W. F., Marin, M. M., & Stewart, L. (2012). Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *Proceedings of the National Academy of Sciences*, *109*(46), 19027-19032.

Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological science*, *11*(3), 188-195.

Trehub, S. E. (2001). Musical predispositions in infancy. *Annals of the New York Academy of Sciences*, *930*(1), 1-16.

Trehub, S. E. (2003). The developmental origins of musicality. *Nature neuroscience*, *6*(7), 669-673.

Trehub, S. E., & Schellenberg, E. G. (1995). Music: Its relevance to infants. *Annals of child development*, *11*(1), 1-24.

Trehub, S. E., & Trainor, L. (1998). Singing to infants: Lullabies and play songs. *Advances in infancy research*, *12*, 43-78.

Trimmer, C. G., & Cuddy, L. L. (2008). Emotional intelligence, not music training, predicts recognition of emotional speech prosody. *Emotion*, *8*(6), 838.

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: the negativity bias in social-emotional development. *Psychological bulletin*, *134*(3), 383.

Van Zonneveld, L., De Sonneville, L., Van Goozen, S., & Swaab, H. (2019). Recognition of facial emotion and affective prosody in children at high risk of criminal behavior. *Journal of the International Neuropsychological Society*, *25*(1), 57-64.

Vanden Bosch der Nederlanden, C. M., Qi, X., Sequeira, S., Seth, P., Grahn, J. A., Joanisse, M. F., & Hannon, E. E. (2022). Developmental changes in the categorization of speech and song. *Developmental science*, e13346.

Venkatraman, A., Edlow, B. L., & Immordino-Yang, M. H. (2017). The brainstem in emotion: a review. *Frontiers in neuroanatomy*, *11*, 15.

Verbeek, P. (1996). Conflict instigation and conflict resolution in preschool children. *Unpublished doctoral dissertation, Emory University, Atlanta*.

Vidas, D., Calligeros, R., Nelson, N. L., & Dingle, G. A. (2020). Development of emotion recognition in popular music and vocal bursts. *Cognition and emotion*, *34*(5), 906-919.

Vidas, D., Dingle, G. A., & Nelson, N. L. (2018). Children's recognition of emotion in music and speech. *Music & Science*, *1*, 2059204318762650.

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, *22*(4), 720-752.

Vigl, J., Talamini, F., Strauss, H., & Zentner, M. (2024). Prosodic discrimination skills mediate the association between musical aptitude and vocal emotion recognition ability. *Scientific reports*, *14*(1), 16462.

Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of cognitive neuroscience*, *22*(12), 2864-2885.

Warrenburg, L. A. (2020). Comparing musical and psychological emotion theories. *Psychomusicology: Music, Mind, and Brain*, *30*(1), 1.

Wechsler, D. (1999). Wechsler abbreviated scale of intelligence. *Psychological Corporation*.

Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., & Scherer, K. R. (2013). On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology*, *4*, 292.

Whitehead, J. C., & Armony, J. L. (2018). Singing in the brain: Neural representation of music and voice as revealed by fMRI. *Human brain mapping*, *39*(12), 4913-4924.

Widen, S. C. (2013). Children's interpretation of facial expressions: The long path from valence-based to specific discrete categories. *Emotion Review*, *5*(1), 72-77.

Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive development*, *23*(2), 291-312.

Williams, K. E., Barrett, M. S., Welch, G. F., Abad, V., & Broughton, M. (2015). Associations between early shared music activities in the home and later child outcomes: Findings from the Longitudinal Study of Australian Children. *Early Childhood Research Quarterly*, *31*, 113-124.

Wolf, M. C., Muijselaar, M. M., Boonstra, A., & de Bree, E. H. (2019). The relationship between reading and listening comprehension: shared and modality-specific components. *Reading and Writing*, *32*(7), 1747-1767.

Woodard, K., Plate, R. C., Morningstar, M., Wood, A., & Pollak, S. D. (2021). Categorization of vocal emotion cues depends on distributions of input. *Affective Science*, *2*, 301-310.

Woodard, K., Zettersten, M., & Pollak, S. D. (2022). The representation of emotion knowledge across development. *Child development*, *93*(3), e237-e250.

Zatorre, R. J., & Baum, S. R. (2012). Musical melody and speech intonation: Singing a different tune.

Zatorre, R. J., & Salimpoor, V. N. (2013). From perception to pleasure: music and its neural substrates. *Proceedings of the National Academy of Sciences*, *110*(supplement_2), 10430-10437.

Zimmer-Gembeck, M. J., Skinner, E. A., Morris, H., & Thomas, R. (2013). Anticipated coping with interpersonal stressors: Links with the emotional reactions of sadness, anger, and fear. *The Journal of Early Adolescence*, *33*(5), 684-709.

Zupan, B. (2015). Recognition of high and low intensity facial and vocal expressions of emotion by children and adults.

Appendices

Appendix A: Stimuli Normalisation

Instrumental pieces were normalised. Additional pieces (those not from the main stimuli set) were re-recorded by a professional musician using the same recording parameters as employed within the main set of stimuli from Grimaud & Eerola's (2021) study, and peak volume normalised to match the average peak volume for its given emotion category. Then, the peak volume of the whole set of instrumental music stimuli was matched to that of the singing and prosody stimuli (-3 dBFS). This allowed for between-emotion differences in loudness within modalities, and between-modality differences within emotions, while ensuring that stimuli would not be uncomfortable for participants. Instrumental pieces were shortened to between 4-7 seconds to align more closely with vocal and singing stimuli, while ensuring musical structure was maintained. Music files were converted to monophonic files with a 48khz sample rate and a bit depth of 16-bit to match vocal stimuli.

Appendix B: Acoustic Feature Extraction and Raw Acoustic Feature Levels for Instrumental Music, Singing, and Vocal Prosody Stimuli

Features were selected based on theoretical significance and past research (e.g., Juslin & Laukka, 2003; Eerola et al., 2013; Paquetta et al., 2018; Llie & Thompson, 2006; Gabrielsson & Lindstrom, 2010). These were loudness (mean, range, variability), pitch (mean, range, variability), tempo/speech rate, spectral centroid (brightness), sharpness, and spectral flux. Spectral centroid refers to the center of mass of the audio spectrum, with higher values relating to a higher proportion of energy being at higher frequencies (hence a brighter sound – Mitrović et al., 2010). Sharpness is similar to brightness but measures the proportion of high frequency energy in a stimulus, while spectral flux refers to the rate and magnitude of variation in spectral content within a stimulus (Gingras et al., 2014). Features were then correlated to ensure they were distinct and to avoid issues of multicollinearity in later analyses. This produced a set of six distinct features – mean loudness, loudness variation, mean pitch, pitch variation, tempo/rate, and spectral centroid. Coefficients of variation, rather than standard deviation, were employed as measures of variability due to standard deviation's correlation with mean levels. A discriminant analysis with the acoustic parameters as independent variables, and emotion as the dependent variable, correctly categorised 40% of stimuli for prosody (chance = 20%; Wilks $\lambda = 0.01$; $F(24, 36.10) = 4.08$, $p<.001$), 45% for instrumental music (Wilks $\lambda = 0.01$; $F(24, 36.10) = 3.55$, $p<.001$), and 85% for singing stimuli (Wilks $\lambda < .001$; $F(24, 36.10) = 12.41$, $p<.001$).

Acoustic feature levels other than tempo/speech rate were extracted using openSMILE software (Eyben et al., 2013). This software operates by taking continuous low-level measurements of acoustic features over short intervals and providing summaries including means and coefficients of variation across the length of the given stimulus. For music, tempo represented the stimulus beats per minute (BPM). For singing and prosody

stimuli, speech rate represented stimulus duration (all stimuli were structurally identical in terms of number of syllables, and the start and end of each stimulus was edited to remove dead noise). The raw feature levels are presented in Tables B1-B3 below, as well as averages for each emotion category.

Table B1 – Raw Acoustic Feature Levels for Instrumental Music Stimuli

Emotion	Stim. number	Mean loudness	Loudness variation	Mean F0	F0 variation	Spectral centroid	Tempo (bpm)
Anger	1	1.836	0.250	81.339	0.242	303.942	100
	2	1.963	0.286	171.290	0.612	428.569	112
	3	2.046	0.241	174.564	0.741	440.272	160
	4	1.879	0.400	119.161	0.316	569.582	130
	Average	1.931	0.294	136.588	0.478	435.591	125.5
Fear	1	1.500	0.338	158.600	0.286	453.763	100
	2	1.604	0.267	132.133	0.656	475.463	110
	3	1.305	0.405	167.878	0.645	468.121	100
	4	1.177	0.330	261.652	0.386	516.301	96
	Average	1.397	0.335	180.066	0.493	478.412	101.5
Sadness	1	0.529	0.182	112.086	0.584	304.614	70
	2	0.426	0.509	191.452	0.633	330.414	80
	3	0.850	0.333	229.082	0.605	434.750	70
	4	0.612	0.404	88.452	0.195	327.008	40
	Average	0.604	0.357	155.268	0.504	349.196	65
Calmness	1	0.508	0.445	314.546	0.341	690.865	91
	2	0.610	0.418	212.929	0.401	536.213	100
	3	0.596	0.200	193.593	0.468	371.352	63
	4	0.582	0.328	115.831	0.462	380.646	72
	Average	0.574	0.348	209.225	0.418	494.769	81.5
Happiness	1	0.787	0.530	179.185	0.696	447.582	110
	2	1.190	0.451	286.578	0.505	546.245	120
	3	0.975	0.207	326.747	0.493	511.983	110
	4	1.196	0.221	167.773	0.553	451.865	180
	Average	1.037	0.352	240.071	0.562	489.419	130

Table B2 – Raw Acoustic Feature Levels for Singing Stimuli

Emotion	Stim. number	Mean loudness	Loudness variation	Mean F0	F0 variation	Spectral centroid	Speech rate
Anger	1	1.805	0.618	302.195	0.396	1874.967	3.111
	2	2.064	0.635	316.562	0.396	2652.156	3.688
	3	2.283	0.438	309.443	0.399	1947.292	2.824
	4	2.402	0.631	322.443	0.299	1875.776	2.752
	Average	2.139	0.581	312.661	0.373	2087.548	3.094
Fear	1	0.928	0.413	312.439	0.418	1733.131	3.186
	2	1.278	0.538	320.931	0.352	2261.713	3.776
	3	0.942	0.535	312.192	0.397	1664.048	2.761
	4	1.084	0.494	324.266	0.319	1510.284	2.767

	Average	1.058	0.495	317.457	0.372	1792.294	3.123
Sadness	1	0.430	0.431	334.350	0.323	1223.647	2.208
	2	0.764	0.484	324.419	0.355	1871.599	3.175
	3	0.614	0.554	322.556	0.374	1294.602	2.474
	4	1.111	0.455	344.682	0.261	1376.199	1.902
	Average	0.730	0.481	331.502	0.328	1441.512	2.440
Calmness	1	0.466	0.413	329.809	0.376	1404.182	2.476
	2	0.639	0.504	341.475	0.299	1251.984	1.931
	3	0.893	0.523	343.450	0.276	1278.327	2.121
	4	0.787	0.412	355.424	0.236	1281.595	1.618
	Average	0.696	0.463	341.540	0.297	1304.022	2.037
Happiness	1	0.855	0.439	341.426	0.333	1759.474	2.756
	2	1.699	0.576	336.539	0.344	2509.219	3.274
	3	1.623	0.537	335.495	0.319	1672.323	2.626
	4	1.578	0.423	362.919	0.149	1363.648	2.739
	Average	1.439	0.494	344.095	0.286	1826.166	2.894

Table B3 – Raw Acoustic Feature Levels for Vocal Prosody Stimuli

Emotion	**Stim. number**	**Mean loudness**	**Loudness variation**	**Mean F0**	**F0 variation**	**Spectral centroid**	**Speech rate**
Anger	1	0.953	0.681	197.596	0.587	2059.115	3.311
	2	1.578	0.671	209.636	0.503	2141.469	3.367
	3	2.316	0.641	269.610	0.551	2374.571	4.021
	4	2.096	0.702	290.114	0.484	3197.790	4.263
	Average	1.736	0.674	241.739	0.531	2443.236	3.741
Fear	1	1.480	0.582	457.405	0.341	1810.938	4.566
	2	1.143	0.533	324.420	0.313	1684.036	4.394
	3	1.259	0.446	281.406	0.355	1518.026	3.842
	4	1.807	0.570	370.037	0.441	2388.243	4.733
	Average	1.422	0.533	358.317	0.362	1850.311	4.384
Sadness	1	0.221	0.764	133.969	0.920	2512.372	2.621
	2	0.418	0.521	161.098	0.518	1656.313	3.784
	3	0.377	0.602	140.406	0.594	1664.303	3.953
	4	1.182	0.594	396.507	0.418	2348.847	4.049
	Average	0.549	0.620	207.995	0.612	2045.459	3.601
Calmness	1	0.234	0.486	163.754	0.611	1607.594	3.232
	2	0.193	0.605	107.588	0.867	1567.598	3.500
	3	0.380	0.480	171.870	0.432	1182.761	3.436
	4	0.271	0.402	181.020	0.561	1511.051	3.382
	Average	0.270	0.493	156.058	0.618	1467.251	3.387
Happiness	1	1.103	0.543	298.491	0.423	1633.237	3.898
	2	0.932	0.465	265.865	0.415	1906.765	4.263
	3	2.541	0.559	389.212	0.384	1842.883	4.230
	4	1.107	0.562	243.204	0.496	2297.394	4.522
	Average	1.421	0.532	299.193	0.430	1920.070	4.228

Appendix C: Z-Scored Acoustic Feature Levels for Instrumental Music, Singing, and Vocal Prosody Stimuli

Some analyses adopted standardised z-scored levels of acoustic features. These z-scores were calculated for all acoustic features, within each condition – presented in tables C1-C3 below.

Table C1 – Z-Scored Acoustic Feature Levels for Instrumental Music Stimuli

Emotion	Stim. number	Mean loudness	Loudness variation	Mean F0	F0 variation	Spectral centroid	Tempo (bpm)
Anger	1	1.343	-0.836	-1.457	-1.563	-1.487	-0.022
	2	1.578	-0.493	-0.183	0.758	-0.214	0.350
	3	1.730	-0.916	-0.137	1.571	-0.094	1.838
	4	1.422	0.601	-0.921	-1.101	1.227	0.908
	Average	1.518	-0.411	-0.675	-0.084	-0.142	0.769
Fear	1	0.723	0.008	-0.363	-1.287	0.044	-0.022
	2	0.915	-0.672	-0.738	1.038	0.266	0.288
	3	0.363	0.646	-0.232	0.967	0.191	-0.022
	4	0.126	-0.069	1.096	-0.658	0.683	-0.146
	Average	0.532	-0.022	-0.059	0.055	0.296	0.025
Sadness	1	-1.069	-1.479	-1.021	0.587	-1.480	-0.951
	2	-1.260	1.639	0.102	0.889	-1.217	-0.641
	3	-0.478	-0.039	0.635	0.718	-0.151	-0.951
	4	-0.917	0.634	-1.356	-1.861	-1.252	-1.881
	Average	-0.931	0.189	-0.410	0.083	-1.025	-1.106
Calmness	1	-1.109	1.026	1.845	-0.946	2.467	-0.301
	2	-0.921	0.774	0.406	-0.564	0.886	-0.022
	3	-0.946	-1.306	0.132	-0.143	-0.798	-1.168
	4	-0.973	-0.090	-0.968	-0.185	-0.703	-0.889
	Average	-0.987	0.101	0.354	-0.460	0.463	-0.595
Happiness	1	-0.593	1.841	-0.072	1.287	-0.019	0.288
	2	0.150	1.084	1.449	0.090	0.989	0.598
	3	-0.246	-1.240	2.017	0.014	0.639	0.288
	4	0.161	-1.113	-0.233	0.388	0.024	2.457
	Average	-0.132	0.143	0.790	0.445	0.408	0.908

Table C2 – Z-Scored Acoustic Feature Levels for Singing Stimuli

Emotion	Stim. number	Mean loudness	Loudness variation	Mean F0	F0 variation	Spectral centroid	Speech rate
Anger	1	1.637	1.774	-1.346	0.653	1.010	0.876
	2	1.281	1.455	-1.079	1.354	0.964	0.703
	3	1.496	-1.737	-1.034	0.857	1.324	0.936
	4	1.599	1.671	-1.073	0.694	1.677	0.717
	Average	1.503	0.791	-1.333	0.889	1.243	0.808
Fear	1	0.056	-0.570	-0.715	1.206	0.491	1.057
	2	-0.018	-0.157	-0.666	0.082	0.270	0.822

	3	-0.487	0.381	-0.847	0.813	0.327	0.714
	4	-0.488	0.121	-0.973	0.999	0.122	0.744
	Average	-0.234	-0.056	-0.800	0.775	0.302	0.834
Sadness	1	-0.842	-0.366	0.635	-1.138	-1.374	-1.300
	2	-0.866	-1.053	-0.337	0.171	-0.422	0.008
	3	-0.970	0.799	-0.141	0.392	-0.974	-0.309
	4	-0.446	-0.321	0.150	0.118	-0.448	-0.820
	Average	-0.781	-0.235	0.077	-0.114	-0.804	-0.605
Calmness	1	-0.776	-0.562	0.355	0.171	-0.713	-0.653
	2	-1.073	-0.722	1.275	-1.451	-1.522	-1.676
	3	-0.559	0.127	1.282	-1.435	-1.032	-1.570
	4	-0.959	-0.796	0.741	-0.248	-0.850	-1.334
	Average	-0.842	-0.488	0.913	-0.741	-1.029	-1.308
Happiness	1	-0.075	-0.276	1.071	-0.892	0.587	0.021
	2	0.677	0.476	0.808	-0.157	0.710	0.143
	3	0.520	0.430	0.740	-0.626	0.356	0.230
	4	0.293	-0.675	1.154	-1.564	-0.501	0.693
	Average	0.354	-0.011	0.943	-0.810	0.288	0.272

Table C3 – Z-Scored Acoustic Feature Levels for Vocal Prosody Stimuli

Emotion	**Stim. number**	**Mean loudness**	**Loudness variation**	**Mean F0**	**F0 variation**	**Spectral centroid**	**Speech rate**
Anger	1	0.279	0.626	-0.401	0.049	0.359	-0.291
	2	1.302	1.402	-0.048	-0.097	1.508	-1.088
	3	0.914	1.167	0.194	0.840	1.494	0.424
	4	1.136	1.268	-0.068	0.071	1.422	0.141
	Average	0.908	1.116	-0.081	0.216	1.196	-0.203
Fear	1	1.228	-0.262	1.577	-1.059	-0.303	1.412
	2	0.521	-0.321	1.300	-1.006	-0.462	1.172
	3	-0.112	-1.218	0.314	-1.036	-0.451	-0.185
	4	0.728	0.035	0.831	-0.709	0.066	1.044
	Average	0.591	-0.442	1.005	-0.952	-0.287	0.861
Sadness	1	-1.040	1.376	-0.885	1.543	1.568	-1.228
	2	-0.780	-0.474	-0.618	-0.026	-0.581	-0.171
	3	-0.969	0.691	-1.117	1.250	-0.119	0.192
	4	-0.157	0.258	1.128	-1.121	0.000	-0.271
	Average	-0.736	0.463	-0.373	0.412	0.217	-0.369
Calmness	1	-1.016	-1.127	-0.658	0.155	-0.846	-0.399
	2	-1.185	0.570	-1.246	1.646	-0.963	-0.796
	3	-0.965	-0.805	-0.798	-0.301	-1.212	-1.568
	4	-1.444	-1.524	-1.295	1.467	-1.402	-1.552
	Average	-1.153	-0.722	-0.999	0742	-1.106	-1.079
Happiness	1	0.549	-0.614	0.367	-0.688	-0.778	0.505
	2	0.142	-1.176	0.612	-0.517	0.497	0.883
	3	1.132	0.165	1.407	-0.753	0.287	1.136
	4	-0.263	-0.036	-0.596	0.292	-0.086	0.638
	Average	0.390	-0.415	0.448	-0.417	0.020	0.791

Appendix D: Jeffrey's (1998) Specifications for Interpreting Bayes Factors

BF_{10}			Interpretation
	>	100	Decisive evidence for H_1
30	-	100	Very strong evidence for H_1
10	-	30	Strong evidence for H_1
3	-	10	Moderate evidence for H_1
1	-	3	Anecdotal evidence for H_1
	1		No evidence
1/3		1	Anecdotal evidence for H_0
1/10		1/3	Moderate evidence for H_0
1/30		1/10	Strong evidence for H_0
1/100		1/30	Very strong evidence for H_0
	<	1/100	Decisive evidence for H_0

Appendix E: Assumption Tests and Reporting

Frequentist

Residual normality and homogeneity of variance was screened for all linear models, including LMMs and multiple linear regressions, via visual inspection of Q-Q plots, histograms, and scatter plots of predicted values by residuals. For LMMs, random effect residuals were also checked.

For GLMMs, assumptions were assessed following the guidelines of Hartig (2022) using the *DHARMa* package in R. This involved calculating simulated residuals that transform them onto a standardised scale, and using these to assess residual normality, variance, and overdispersion. This was done for the model as a whole and for individual random effects (Bono et al., 2021).

Correlations were run with the *stats* package (R Core Team, 2018). LMM analyses were conducted with the *lmertest* package (Kutsenova et al., 2020), and GLMMs with *lme4* (Bates et al., 2014). Both were fit with maximum likelihood estimation. Fixed effects for these analyses were further explored via the *afex* package (Singmann et al., 2020). For LMMs, *F*-values and *p*-values for fixed effects were reported, calculated using Satterwhaite's method to approximate degrees of freedom (Kuznetsova et al., 2017). For GLMMs, fixed effects were reported in terms of type 3 Likelihood Ratio Chi-square statistics and associated *p*-values. Random effects for each model were estimated through comparing the final model to a series of reduced models that removed each random effect via Likelihood Ratio Tests. All parametric post-hoc pairwise comparisons were adjusted using Tukey's *HSD* correction within the *emmeans* package (Lenth, 2021). Estimated marginal means were reported for all parametric analyses with countable outcomes, and odds ratios were reported for those with categorical outcomes. Standard errors were also reported for parametric tests. Alpha value was set at 0.05.

Bayesian Tests Reporting and Priors

For Bayesian analyses, all priors were default priors, as advised by Rouder and Morey (2012). For LMMs with categorical predictors, these were selected following Rouder et al. (2012) - Cauchy prior distributions with scale set to $r = 0.5$ for fixed factors, $r = 1$ for random factors, and $r = 0.354$ for covariates (continuous predictors). For linear regression models, these were Jeffrey-Zellner-Siow priors – a special case of a Cauchy prior with scale set to 0.354. For correlations, the prior was a stretched beta prior width of 1. For GLMMs, default priors were selected following Oberauer (2019, 2023) – Cauchy prior distributions with scale set to 0.35 for fixed effects, and uninformative Gamma prior distribution with a mean of 1 and *SD* of 0.04 for random effects. In each case, BFs were only produced for fixed effects. One exception to the inclusion of BFs was the GLMM analysis of the influence of stimulus arousal and valence on emotion selection. This was due to instability in BF estimates (different estimates with each computation), possibly due to model complexity (multiple continuous by continuous interactions, by-participant random slopes) relative to other models. This was the case even at high numbers of model iterations (50,000).

Bayesian models equivalent in fixed and random effects structure to frequentist models were fit with the *brms* package (Burkner, 2017). BFs were reported at the level of main/interaction effects, and pairwise comparisons between factor levels. For fixed main and interaction effects, the full model was compared to models with systematically reduced fixed effects structures (Oberauer, 2022; Van Doorn et al., 2021). This was done using the *BayesFactor* package (Morey et al., 2018). For post-hoc comparisons, to maintain the effects of random factors and compare marginal means, BFs were estimated via Savage-Dickey density ratios. This involved comparing the estimated model to a model within which the comparison of focus was restricted to the point-null (zero). This was done using the *BayesTestR* (Makowski et al., 2019) and *emmeans* (Lenth, 2021) R packages.

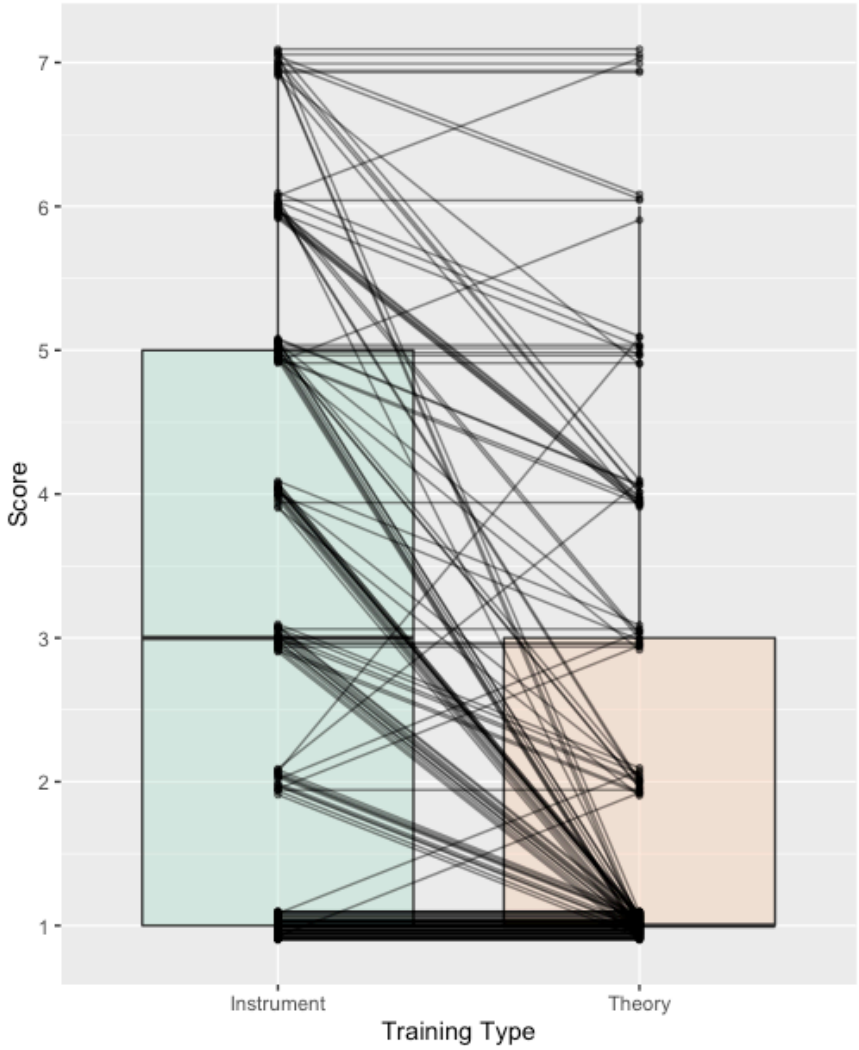
Appendix F: Mixed Model Selection Procedure

All mixed models had the maximal random effect structure afforded by the design (Barr et al., 2013). This included random intercepts for participant and, where relevant, stimuli, and random by-participant/by-stimuli slopes for each categorical predictor (as well as interactions between these predictors). If models failed to converge or had issues with singularity (an overly complex random effect structure), random slopes were systematically removed from the model, starting at the highest level (i.e., slopes for interactions). In line with Matuschek et al. (2017), those models that converged successfully were compared with a Likelihood Ratio Test (LRT) via a backwards-selection heuristic (starting with the most complex model and reducing complexity). The final model was chosen if a further reduction in complexity would reduce model fit with an LRT alpha level of 0.2. Bayesian models were fit with a matching fixed and random effects structure to allow comparable results.

Appendix G: Chapter 2 Music Training Score Distribution and Composite Measure

Figure G1 shows the distribution of years of instrumental and theory music training among participants.

Figure G1 – Distribution of Musical Instrument and Theory Training



Note. Score = years of formal training (Instrument, 1 = 0, 2 = 0.5, 3 = 1, 4 = 2, 5 = 3-5, 6 = 6-9, 7 = 10 or more; theory, 1 = 0, 2 = 0.5, 3 = 1, 4 = 2, 5 = 3, 6 = 4-6, 7 = 7 or more). Lines connect individual participants – more lines = more connections between training scores. $N = 159$.

Instrument training was widely distributed. In general, participants with musical training had more instrument than theory training. A spearman's correlation indicated that

years of instrumental and theory training ($r(157) = 0.67$, $p<.001$, $BF_{10}>100$) were highly related. Accordingly, the two constructs were combined to give a composite music training score per participant for use in analyses.

Appendix H: Chapter 2 Emotion Recognition Accuracy (%) and Confusion Patterns

Table H1 - Instrumental music

		Selected Emotion					
		Anger	**Fear**	**Sadness**	**Calmness**	**Happiness**	**None**
Target emotion	**Anger**	**59.9 (24.2)**	26.7 (22.2)	2.2 (7.1)	1.1 (5.9)	4.4 (9.9)	5.7 (13.2)
	Fear	18.9 (21.7)	**69.7 (25.9)**	3.1 (9.2)	0.9 (4.8)	2.4 (7.9)	5.0 (13.1)
	Sadness	0.9 (5.5)	3.1 (8.3)	**82.9 (22.5)**	11.2 (17.9)	0.8 (5.2)	1.1 (5.1)
	Calmness	0.8 (6.5)	1.26 (6.2)	20.0 (21.9)	**63.2 (27.5)**	13.1 (18.2)	1.7 (6.4)
	Happiness	1.4 (5.8)	0.8 (4.4)	0.8 (6.5)	2.8 (8.4)	**92.5 (15.3)**	1.7 (7.5)

Note. Mean (standard deviation). $N = 159$. **Bold** diagonal = correct recognition.

Table H2 - Singing

		Selected Emotion					
		Anger	**Fear**	**Sadness**	**Calmness**	**Happiness**	**None**
Target emotion	**Anger**	**74.8 (23.6)**	10.5 (15.5)	1.6 (6.7)	0.8 (4.4)	6.8 (13.4)	5.5 (13.7)
	Fear	8.2 (12.7)	**53.5 (23.1)**	18.2 (17.9)	4.2 (10.2)	7.2 (13.0)	8.6 (13.5)
	Sadness	1.3 (5.5)	12.7 (17.3)	**49.7 (22.3)**	15.6 (18.6)	13.4 (13.1)	7.4 (13.9)
	Calmness	0.3 (2.8)	4.1 (9.3)	31.0 (23.8)	**51.7 (25.6)**	9.4 (14.5)	3.5 (9.9)
	Happiness	3.1 (9.6)	12.9 (17.5)	13.4 (17.1)	11.6 (15.9)	**49.7 (27.1)**	9.3 (16.8)

Note. Mean (standard deviation). **Bold** diagonal = correct recognition.

Table H3 - Prosody

		Selected Emotion					
		Anger	**Fear**	**Sadness**	**Calmness**	**Happiness**	**None**
Target emotion	**Anger**	**94.3 (11.9)**	1.7 (7.0)	0.6 (4.8)	0.6 (3.9)	0.9 (4.8)	1.7 (7.0)
	Fear	5.2 (11.6)	**73.0 (22.7)**	12.9 (14.6)	0.20 (2.0)	7.5 (14.5)	1.3 (5.5)
	Sadness	2.7 (8.2)	31.0 (19.8)	**56.4 (21.5)**	4.4 (11.1)	2.4 (7.3)	3.1 (8.8)
	Calmness	0.8 (4.4)	1.6 (6.7)	18.4 (21.7)	**68.7 (25.8)**	6.0 (11.1)	4.6 (12.5)
	Happiness	7.5 (12.2)	19.8 (19.5)	6.8 (11.8)	3.3 (9.4)	**54.6 (26.8)**	8.0 (14.2)

Note. Mean (standard deviation). **Bold** diagonal = correct recognition.

Appendix I: Adult Valence and Arousal Ratings by Emotion and Condition

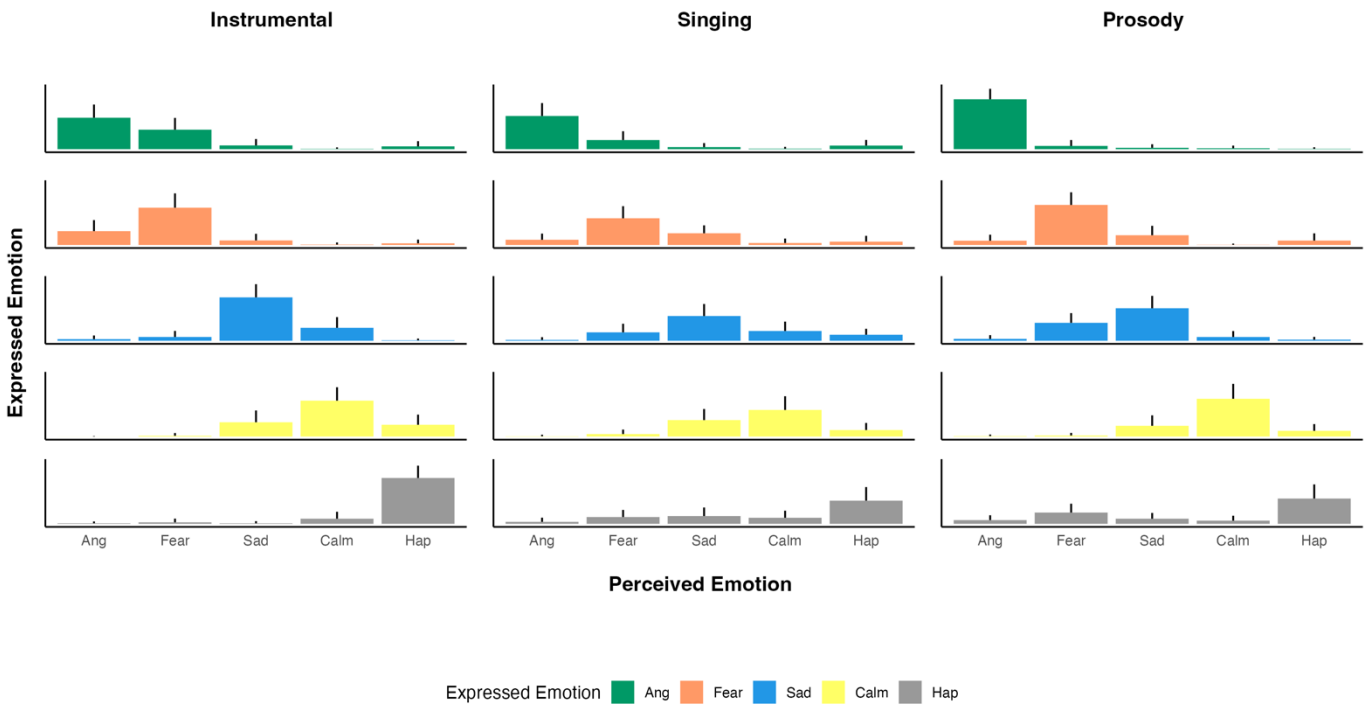
	Ratings					
	Instrumental		Singing		Prosody	
Emotion	**Valence**	**Arousal**	**Valence**	**Arousal**	**Valence**	**Arousal**
Anger	0.23	0.87	0.22	0.85	0.11	0.83
	(0.21)	(0.17)	(0.23)	(0.18)	(0.14)	(0.20)
Fear	0.23	0.78	0.24	0.61	0.24	0.73
	(0.19)	(0.20)	(0.23)	(0.25)	(0.26)	(0.23)
Sadness	0.17	0.16	0.25	0.24	0.15	0.28
	(0.17)	(0.16)	(0.24)	(0.24)	(0.18)	(0.30)
Calmness	0.54	0.20	0.44	0.18	0.48	0.11
	(0.28)	(0.18)	(0.27)	(0.17)	(0.25)	(0.14)
Happiness	0.83	0.75	0.64	0.65	0.61	0.69
	(0.22)	(0.21)	(0.29)	(0.26)	(0.31)	(0.23)

Note. Mean (standard deviation). Scores range from 0-1. N =157.

Appendix J: Chapter 2 Patterns of Emotion Scale Ratings for Each Condition

Scale ratings were almost identical to emotion recognition patterns, suggesting participants generally perceived stimuli as unambiguously expressive of a single emotion, despite errors in this regard. Figure J1 outlines scale rating patterns.

Figure J1 – Patterns of Average Scale Ratings for Each Expressed Emotion and Condition



Note. Error bars = standard deviation. Min = 1; Max = 5. $N = 159$.

For instrumental music, there was a significant effect of scale on ratings for all expressed emotions (χ^2 (4) = 485.45 for anger, 469.38 for fear, 491.04 for sadness, 476 for calmness, 481.84 for happiness, *ps*<.001). The same was true for singing (χ^2 (4) = 448.65 for anger, 403.35 for fear, 343.60 for sadness, 433.16 for calmness, 243.69 for happiness,

ps<.001) and for prosody (χ^2 (4) = 489.22 for anger, 436.70 for fear, 487.76 for sadness, 449.26 for calmness, 302.74 for happiness). Post-hoc tests signalled that the expressed emotion was rated higher on its matched scale than any other scales across conditions and emotions (*ps*<.01).

Appendix K: Chapter 3 Emotion Recognition Accuracy (%) and Confusion Patterns by Condition, Emotion, and Age Group

Table K1 – Instrumental Recognition Accuracy (%) by Emotion and Age Group

Target Emotion	Selected Emotion				
	Anger	**Fear**	**Sadness**	**Calmness**	**Happiness**
4-5 years					
Anger	**61.11 (34.9)**	21.3 (29.17)	6.48 (16.4)	0.93 (4.81)	10.19 (15.9)
Fear	37.96 (30.52)	**24.07 (26.39)**	7.41 (13.54)	5.56 (10.59)	25 (31.01)
Sadness	3.7 (15.04)	12.96 (24.39)	**39.81 (31.21)**	34.26 (33.36)	9.26 (18.54)
Calmness	1.85 (6.67)	2.78 (8.01)	27.78 (26.25)	**48.15 (31.72)**	19.44 (22.29)
Happiness	10.19 (19.93)	6.48 (14.86)	2.78 (8.01)	16.67 (20.8)	**63.89 (35.58)**
6-7 years					
Anger	**61.03 (37.03)**	35.29 (35.42)	0.74 (4.29)	0 (0)	2.94 (8.18)
Fear	35.29 (30.85)	**50 (31.98)**	3.68 (10.89)	3.68 (12.51)	7.35 (13.1)
Sadness	0.74 (4.29)	10.29 (20.52)	**73.53 (32.53)**	13.24 (19.69)	2.21 (7.2)
Calmness	0.74 (4.29)	5.88 (15.15)	28.68 (25.47)	**55.88 (30.81)**	8.82 (16.15)
Happiness	5.15 (18.24)	2.21 (9.47)	3.68 (10.89)	5.88 (10.76)	**83.09 (25.17)**
8-9 years					

Target Emotion	Selected Emotion				
	Anger	**Fear**	**Sadness**	**Calmness**	**Happiness**
Anger	**67.65 (29.85)**	27.21 (28.45)	2.21 (9.47)	0 (0)	2.94 (10.23)
Fear	30.88 (30.19)	**52.21 (29.75)**	6.62 (14.18)	2.94 (10.23)	7.35 (13.1)
Sadness	0 (0)	10.29 (15.22)	**70.59 (26.45)**	18.38 (22.45)	0.74 (4.29)
Calmness	0.74 (4.29)	0 (0)	31.62 (28.41)	**65.44 (27.53)**	2.21 (7.2)
Happiness	0.74 (4.29)	2.21 (9.47)	1.47 (5.97)	8.09 (18.17)	**87.5 (24.81)**

Note. Mean (standard deviation). **Bold** diagonal = correct recognition.

Table K2 – Singing Recognition Accuracy (%) by Emotion and Age Group

Target Emotion	Selected Emotion				
	Anger	**Fear**	**Sadness**	**Calmness**	**Happiness**
4-5 years					
Anger	**65.74 (34.07)**	16.67 (24.02)	7.41 (13.54)	0.93 (4.81)	9.26 (15.73)
Fear	12.04 (14.5)	**28.7 (20.46)**	28.7 (25.67)	12.96 (16.07)	17.59 (23.83)
Sadness	3.7 (11.4)	11.11 (14.43)	**34.26 (30.34)**	35.19 (24.28)	15.74 (18.54)
Calmness	2.78 (10.59)	9.26 (15.73)	24.07 (21.35)	**53.7 (28.34)**	10.19 (17.35)
Happiness	16.67 (21.93)	14.81 (17.35)	25.93 (22.45)	20.37 (18.39)	**22.22 (22.29)**
6-7 years					
Anger	**78.68 (23.94)**	13.24 (14.08)	2.94 (8.18)	0 (0)	5.15 (10.26)

Target Emotion	Selected Emotion				
	Anger	**Fear**	**Sadness**	**Calmness**	**Happiness**
Fear	16.91 (19.19)	**38.24 (26.28)**	26.47 (20.36)	10.29 (17.54)	8.09 (13.37)
Sadness	4.41 (9.67)	18.38 (18.78)	**35.29 (22.29)**	31.62 (26.33)	10.29 (16.42)
Calmness	0 (0)	8.82 (13.6)	22.06 (21.11)	**59.56 (28.21)**	9.56 (15.09)
Happiness	9.56 (15.09)	16.18 (16.15)	23.53 (18.4)	22.06 (17.15)	**28.68 (22.3)**
8-9 years					
Anger	**75.74 (27.17)**	10.29 (17.54)	3.68 (10.89)	2.21 (9.47)	8.09 (13.37)
Fear	12.5 (16.57)	**47.79 (29.75)**	22.06 (21.11)	8.09 (17.1)	9.56 (12.33)
Sadness	0.74 (4.29)	17.65 (19.97)	**39.71 (24.71)**	32.35 (19.97)	9.56 (12.33)
Calmness	0.74 (4.29)	10.29 (16.42)	25 (22.19)	**55.15 (21.12)**	8.82 (13.6)
Happiness	8.82 (13.6)	12.5 (15.39)	18.38 (17.74)	16.91 (17.1)	**43.38 (24.08)**

Note. Mean (standard deviation). **Bold** diagonal = correct recognition.

Table K3 – Prosody Recognition Accuracy (%) by Emotion and Age Group

Target Emotion	Selected Emotion				
	Anger	**Fear**	**Sadness**	**Calmness**	**Happiness**
4-5 years					
Anger	**81.48 (30.69)**	7.41 (16.72)	4.63 (12.08)	2.78 (8.01)	3.7 (11.4)
Fear	16.67 (18.34)	**47.22 (30.49)**	24.07 (28.15)	2.78 (8.01)	9.26 (15.73)
Sadness	4.63 (9.9)	15.74 (17.19)	**41.67 (31.77)**	28.7 (25.67)	9.26 (15.73)
Calmness	5.56 (17.45)	3.7 (9.05)	12.04 (24.39)	**69.44 (35.58)**	9.26 (15.73)
Happiness	31.48 (26.49)	21.3 (20.46)	16.67 (18.34)	8.33 (13.87)	**22.22 (23.34)**
6-7 years					
Anger	**94.12 (12.4)**	4.41 (11.47)	1.47 (5.97)	0 (0)	0 (0)
Fear	12.5 (15.39)	**60.29 (28.94)**	25 (23.84)	0.74 (4.29)	1.47 (5.97)
Sadness	1.47 (5.97)	25 (19.46)	**49.26 (22.6)**	22.06 (19.23)	2.21 (7.2)
Calmness	0.74 (4.29)	2.94 (10.23)	10.29 (19.58)	**81.62 (25.61)**	4.41 (9.67)
Happiness	18.38 (22.45)	27.21 (20.75)	11.76 (15.37)	11.03 (15.31)	**31.62 (24.85)**
8-9 years					
Anger	**92.65 (11.56)**	2.94 (8.18)	2.94 (8.18)	0.74 (4.29)	0.74 (4.29)
Fear	5.15 (11.96)	**65.44 (28.21)**	25 (21.32)	0 (0)	4.41 (11.47)

Target Emotion	Selected Emotion				
	Anger	**Fear**	**Sadness**	**Calmness**	**Happiness**
Sadness	4.41 (11.47)	20.59 (21.73)	**58.82 (24.53)**	14.71 (19.58)	1.47 (5.97)
Calmness	2.21 (7.2)	2.21 (7.2)	9.56 (20.43)	**81.62 (24.85)**	4.41 (9.67)
Happiness	14.71 (19.58)	30.88 (22.24)	11.03 (16.5)	5.88 (15.15)	**37.5 (25.56)**

Note. Mean (standard deviation). **Bold** diagonal = correct recognition.

Appendix L: Stimuli Presentation Order A and Order B

	Order A	Order B
1.	SAD	SCARED
2.	SCARED	CALM
3.	ANGRY	ANGRY
4.	CALM	SCARED
5.	HAPPY	CALM
6.	ANGRY	HAPPY
7.	SCARED	SCARED
8.	SAD	SAD
9.	HAPPY	HAPPY
10.	CALM	SAD
11.	SCARED	ANGRY
12.	HAPPY	SCARED
13.	SAD	ANGRY
14.	HAPPY	CALM
15.	CALM	ANGRY
16.	ANGRY	CALM
17.	CALM	HAPPY
18.	SAD	SAD
19.	SCARED	HAPPY
20.	ANGRY	SAD
21.	CALM	HAPPY
22.	HAPPY	SAD
23.	ANGRY	CALM
24.	SAD	ANGRY
25.	SCARED	SAD
26.	ANGRY	HAPPY
27.	SAD	ANGRY
28.	SCARED	CALM
29.	CALM	HAPPY
30.	HAPPY	SCARED
31.	CALM	SAD
32.	ANGRY	CALM
33.	SCARED	ANGRY
34.	SAD	HAPPY
35.	HAPPY	SCARED
36.	ANGRY	CALM
37.	SAD	SAD
38.	CALM	SCARED
39.	SCARED	ANGRY
40.	HAPPY	SCARED

Appendix M: Chapter 4 Emotion Language Comprehension Task Testing Protocol and Scoring Sheet

Emotion Vocabulary Assessment – Testing and Scoring Guidelines

Source: *Nook et al. (2020) – Charting the development of emotion comprehension and abstraction from childhood to adulthood using observer-rated and linguistic measures.*

Materials. 15 laminated index cards with emotions written on them, scoring manual, scoring sheet, and audio recorder.

Example emotion: Surprised

Emotion set 1: Angry, calm, happy, sad, scared

Emotion set 2: Content, disappointed, embarrassed, excited, frustrated, jealous, nervous, proud, relaxed, worried

General recommendations. Make sure that each set of emotion cards has been shuffled multiple times prior to starting the task so that emotion order is random. Avoid providing corrective or praising feedback in this task. Attempt to keep them motivated but not bias their answers. Use encouraging but neutral feedback. For example, you might say things like "ok", "let's try another", or "you're doing good work" rather than "that's correct". Try not to cut them off after they provide a suitable response, as they should share whatever they think about each emotion.

Procedure. Begin by providing them with the following instructions: "This game is called the Emotion Word Game. In this game, I'm going to say a word, and I want you to do your best to tell me what the word means. I'll show you an example first." Place 'surprised' card in front of participant and say "surprised is when someone feels shocked or amazed because something happens that they don't expect. You might be surprised when opening a present that you didn't expect to get. OK, are you ready to start?" Start the audio recorder. On each trial, pull a card from the deck of emotion terms (or ask the participant to pull the card), show it to them, and say: "What does _____ mean?". Follow the scoring guidelines below for each trial and probe them if their responses do not initially earn 2 points. Complete emotion set 1 first. If participants a) cannot complete the trials due to difficulties understanding or expressing answers, b) have scored 0 (no understanding of any of the set 1 words), stop here. Otherwise, continue to set 2, and complete all emotion cards.

Overall scoring guidelines.

Two-point answers:

a. A reasonable description of the definition, even if the definition is not *exact* (see example definitions below)

b. An example situation that would reasonably evoke the emotion in question and *not other emotions* (e.g., "I felt sad when my pet died" for sad),

c. A synonym (even if colloquial) or closely related emotional term (e.g., "pissed off" or "mad" for angry) *if also put into a suitable context* (i.e., a situation or reasonable description).

***One-point answers*:**

a. A response that is of the correct valence but overly vague. No features are provided to distinguish this emotion from other similar emotions (e.g., "good," "bad," "positive," "negative," etc.). This includes responses that involve situations that could give rise to several emotions of a similar valence as the target emotion.

b. A synonym without any further elaboration.

***Zero-points*:**

When the participant clearly does not know what the emotion word means, even after probing. These responses include:

a. An incorrect definition of the term

b. An example of a situation that is unlikely to evoke the emotion in question (e.g., "I feel happy when my pet dies")

c. The response "I don't know."

Situation understanding and definition strategy:

Participants received either 1 or 0 points for their situation understanding (based on their responses after probing). This was to examine if children understood how emotions are experienced in context and look at any individual differences here, as although situational definitional strategies fall throughout childhood, it may be that children with emotional difficulties show a less natural and/or early grasp of situation-based emotion comprehension.

Further coding was done for the unassisted definitional strategy used. This included reference to situations (before probing) and the use of general definitions (i.e., a description of the emotion word not tied to a specific context or situation).

Each of these aspects was only scored for 2-point trials, to ensure that patterns reflected a) full situation understanding and b) successful use of strategies.

Queries.

Query participants to assess their understanding of the emotion term. Score their response based on their responses to queries (i.e., they can earn 1 or 2 points for the emotion if later responses clarify their understanding of the emotion term).

If they scored 1 point:

1. If overly vague definition or synonym provided: “Can you tell me more about [emotion]?”
2. If situation is non-specific (e.g., “I felt angry when my sister got something that I wanted,” which could evoke jealousy or anger): “Can you tell me more about that situation?” or “can you give me another situation where someone might be [emotion]?”

If they scored 0 points:

3. “What other feelings might be like [emotion]?” or “what other words do people use to describe [emotion]?”

If participant did NOT provide an example situation (even if they scored 2 points):

4. “Can you tell me when someone might be [emotion]?” or “what are some things that might make someone [emotion]?”.

However, use these latter forms of encouragement (0-point probes for synonyms/situations) sparingly (a maximum of 1 synonym and 1 situation probe). One goal of the assessment is to measure how much people generate these types of strategies, so only use them if probing suggests they are not able to spontaneously generate a full definition.

Additional scoring guidelines:

a. Try not to query a query (e.g., if a participant gives a synonym after a query, do not query what this synonym means. Instead, you can ask a different question about the target emotion, such as “when might someone feel [emotion]?)
b. Poor articulation or poor grammar do NOT affect score
c. Extraneous comments do NOT affect score
d. Response CAN be “spoiled” and scored 0 if examinee adds information that indicates he or she really does not understand the meaning of the word

Definition and synonym guidelines.

Definitions below are based on: Merriam-Webster Dictionary for Adults, Merriam-Webster Dictionary for Children, Oxford Dictionary, and Google Dictionary. Synonyms are taken from a variety of sources. These definitions are only to serve as a rough guide in your scoring.

Angry. Definition: strong feeling of being upset, annoyed, displeased, or hostile. Synonyms: irate, mad, annoyed, cross, vexed, irritated, indignant, irked, furious, enraged, infuriated, in a temper, displeasure, fury, aggravated, livid; ticked off, pissed off; losing one’s temper.

Calm. Definition: in a quiet and peaceful state or condition; not feeling or showing nervousness, anger or other emotions. Synonyms: serene, tranquil, relaxed, unruffled, unperturbed, unflustered, untroubled.

Content. Definition: pleased and satisfied with what one has or is. Synonyms: delighted, glad, gratified, satisfied, pleased, fulfilled, tranquil, at ease.

Disappointed. Definition: sad, unhappy, or displeased because someone or something has failed to fulfil one's hopes or expectations. Synonyms: upset, saddened, let down, cast down, disheartened, downhearted, downcast, depressed, dispirited, discouraged, despondent, dismayed, distressed.

Embarrassed. Definition: confused and foolish in front of other people; self- consciousness, shame, or awkwardness. Synonyms: mortified, red-faced, blushing, abashed, shamed, ashamed, humiliated, awkward, self-conscious, uncomfortable...

Excited. Definition: eager enthusiasm and interest. Synonyms: thrilled, exhilarated, animated, enlivened, electrified.

Frustrated. Definition: feeling discouragement, anger, and annoyance because of unresolved problems or unfulfilled goals, desires, or needs. Synonyms: disappointed, disenchanted, unfulfilled, disillusioned, dejected, displeased, vexed, irritated, infuriated, discouraged.

Happy. Definition: pleasure and enjoyment because of life, situation, etc.; contentment. Synonyms: cheerful, cheery, merry, joyful, jovial, jolly, jocular, gleeful, delighted, untroubled, smiling, beaming, grinning, in good spirits, in a good mood, lighthearted, pleased, content, satisfied, gratified, sunny, joyous.

Jealous. Definition: intolerant of rivalry or unfaithfulness; envy of someone or their achievements and advantages. Synonyms: envious, covetous, desirous.

Nervous. Definition: worried and afraid about what might happen; easily agitated or alarmed; tending to be anxious; highly strung. Synonyms: high-strung, anxious, edgy, tense, excitable, jumpy, skittish, brittle, neurotic.

Proud. Definition: deep pleasure, satisfaction, or happiness as a result of one's own achievements, qualities, or possessions or those of someone with whom one is closely associated; attitude or people who think that they are better or more important than others. Synonyms: pleased, glad, happy, delighted, joyful overjoyed, thrilled, satisfied, gratified, content.

Relaxed. Definition: calm and free from stress, worry, or anxiety; free from tension and anxiety; at ease. Synonyms: comfy, cozy, relaxed, content, satisfied, peaceful, resting, easygoing, undisturbed.

Sad. Definition: grief or unhappiness; sorrow. Synonyms: unhappy, sorrowful, dejected, depressed, downcast, miserable, down, blue, down in the dumps, blah.

Scared. Definition: afraid of something; nervous, frightened, fearful. Synonyms: afraid, startled, nervous, fearful, panicky, alarmed, intimidated, terrified, petrified, terrorized, spooked.

***Worried*.** Definition: fear or concern because you think that something bad has happened or could happen; anxious, upset, or troubled about actual or potential problems. Synonyms: anxious, perturbed, troubles, bothered, concerned, upset, distressed, uneasy, agitated, nervous, edgy, tense, keyed up, jumpy, stressed, strung out.

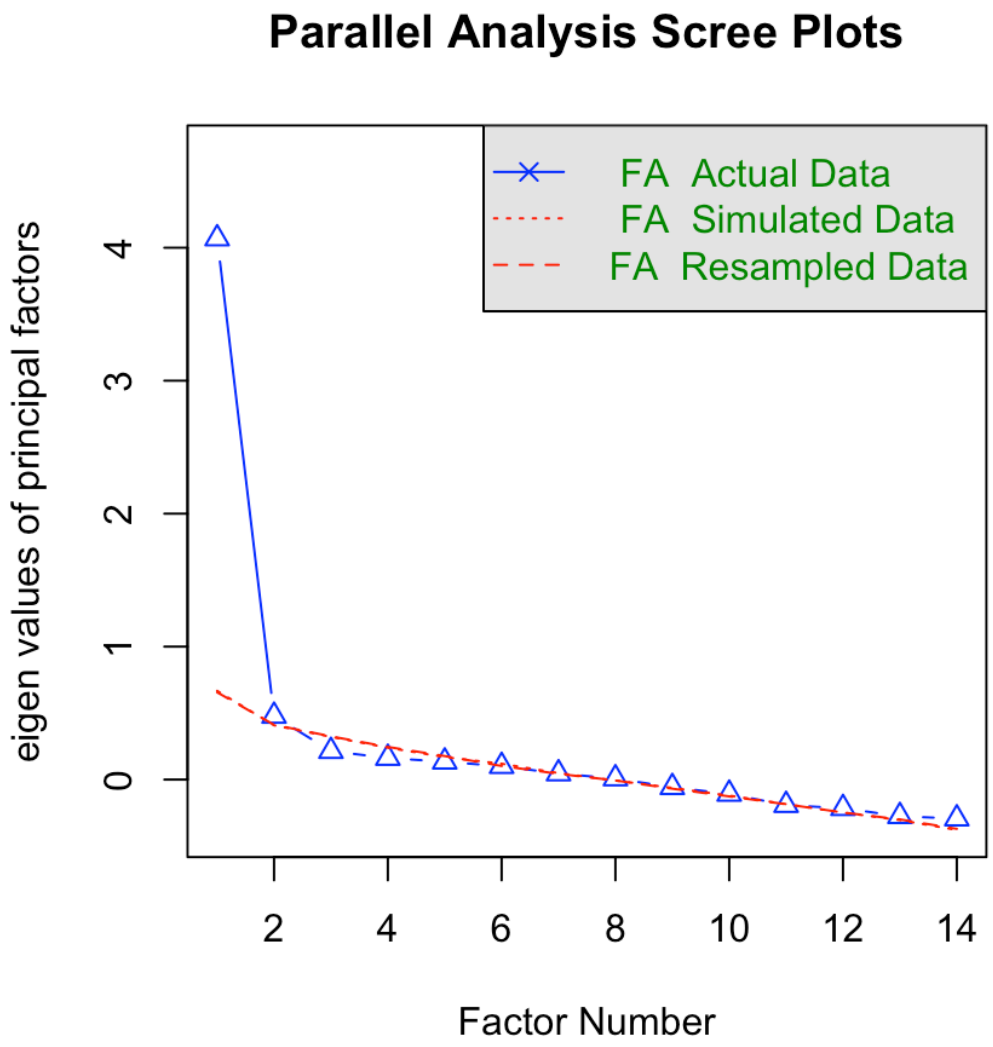
Emotion Vocabulary Assessment – Scoring Sheet

Ppt number:			
Emotion	**0** Does not understand word	**1** Correct valence but vague OR synonym without context	**2** Reasonable definition OR emotion-specific situation OR synonym in context
Angry			
Calm			
Content			
Disappointed			
Embarrassed			
Excited			
Frustrated			
Happy			
Jealous			

Nervous			
Proud			
Relaxed			
Sad			
Scared			
Worried			

Total score:

Appendix N: Scree Plot Showing the Weight (Eigen Value) of Each Factor for Emotion Language Comprehension Variable



Appendix O: 1-Factor Solution for Final Emotion Language Comprehension Variable Items - Means, Standard Deviations, Item-Rest Correlations, and Factor Loadings

Item	Mean	SD	Item-rest	Factor loading
Angry	1.63	0.57	0.41	0.55
Calm	1.61	0.67	0.44	0.60
Disappointed	1.42	0.73	0.57	0.70
Embarrassed	0.90	0.93	0.65	0.86
Excited	1.83	0.50	0.25	0.40
Frustrated	0.93	0.82	0.54	0.67
Happy	1.87	0.42	0.36	0.61
Jealous	1.23	0.94	0.60	0.76
Nervous	1.46	0.72	0.66	0.83
Proud	1.65	0.63	0.55	0.73
Relaxed	1.79	0.53	0.48	0.72
Sad	1.72	0.50	0.38	0.52
Scared	1.91	0.32	0.25	0.43
Worried	1.46	0.72	0.58	0.72

Note. Scored from 0-2. $N = 179$.

Appendix P: Chapter 4 Bivariate Correlations for Referred and Typically Developing Samples Independently

Table P1 – Referred Sample Bivariate Correlations Between Emotion Recognition Accuracy, Emotion Language Comprehension, Receptive Vocabulary (BPVS), Age, Sex, and Years of Music Training

Variable	1	2	3	4	5	6	7	8
1. Accuracy Total	-	-	-	-	-	-	-	-
2. Accuracy Instrumental	-	-	-	-	-	-	-	-
3. Accuracy Singing	-	.44*** (>100)	-	-	-	-	-	-
4. Accuracy Prosody	-	.58*** (>100)	.53*** (>100)	-	-	-	-	-
5. Emotion Comprehension	.54*** (>100)	.30** (>100)	.45*** (>100)	.54*** (>100)	-	-	-	-
6. BPVS	.48*** (>100)	.36*** (46.20)	.33** (18.76)	47*** (>100)	.65*** (>100)	-	-	-
7. Age	.23* (1.69)	.09 (0.34)	.23* (1.78)	.23* (1.98)	.54*** (>100)	.49*** (>100)	-	-
8. Music Training (Years)	.06 (0.30)	-.01 (0.26)	.13 (0.46)	.07 (0.31)	.34** (8.49)	-.01 (0.26)	.12 (0.43)	-
9. Sex	-.11 (0.41)	-.02 (0.26)	-.24* (2.06)	-.11 (0.47)	-.18 (0.69)	.03 (0.26)	-.02 (0.26)	.04 (0.27)

Note. Coefficient (BF_{10}). Spearman's rho correlations involving music training and age. Biserial correlations involving sex. Pearson's correlations between all other variables. *df* = 62 between variables and emotion language comprehension; 78 between all other variables. All variable scores are raw (unadjusted). BPVS = British Vocabulary Picture Scale. $p<.05$*, $p<.01$**, $p<.001$***.

Table P2 – TD Sample Bivariate Correlations Between Emotion Recognition Accuracy, Emotion Language Comprehension, Receptive Vocabulary (BPVS), Age, Sex, and Years of Music Training

Variable	1	2	3	4	5	6	7	8
1. Accuracy Total	-	-	-	-	-	-	-	-
2. Accuracy Instrumental	-	-	-	-	-	-	-	-
3. Accuracy Singing	-	.44*** (>100)	-	-	-	-	-	-
4. Accuracy Prosody	-	.46*** (>100)	.37** (16.59)	-	-	-	-	-
5. Emotion Comprehension	.59*** (>100)	.54** (>100)	.39** (26.91)	.44*** (86.34)	-	-	-	-
6. BPVS	.43*** (79.62)	.50*** (>100)	.16 (0.63)	29* (3.68)	.70*** (>100)	-	-	-
7. Age	.54*** (>100)	.51*** (>100)	.32* (5.93)	.30* (4.06)	.72*** (>100)	.63*** (>100)	-	-
8. Music Training (Years)	.10 (0.39)	.18 (0.69)	-.10 (0.38)	.02 (0.30)	.17 (0.21)	.16 (0.56)	-.02 (0.30)	-
9. Sex	-.22 (1.16)	-.29* (2.99)	-.24* (2.06)	-.06 (0.32)	-.21 (0.98)	-.25* (1.62)	-.22 (1.07)	-.10 (0.38)

Note. Coefficient (BF_{10}). Spearman's rho correlations involving music training and age. Biserial correlations involving sex. Pearson's correlations between all other variables. *df* = 59 between variables and emotion language comprehension; 60 between all other variables. All variable scores are raw (unadjusted). BPVS = British Vocabulary Picture Scale. $p<.05$*, $p<.01$**, $p<.001$***.

Appendix Q: Chapter 4 Partial Correlations Controlling for Age, for Referred and Typically Developing Samples Independently

Table Q1 – Referred Sample Partial Correlations Between Emotion Recognition Accuracy, Emotion Language Comprehension, and Receptive Vocabulary (BPVS), Controlling for Age

Variable	1	2	3	4	5
1. Accuracy Total	-	-	-	-	-
2. Accuracy Instrumental	-	-	-	-	-
3. Accuracy Singing	-	.44*** (>100)	-	-	-
4. Accuracy Prosody	-	.57*** (>100)	.51*** (>100)	-	-
5. Emotion Comprehension	.48*** (>100)	.30** (>100)	.36*** (>100)	.47*** (>100)	-
6. BPVS	.44*** (>100)	.37*** (36.39)	.26* (1.80)	.43*** (>100)	.51*** (>100)

Note. Pearson's correlations. *df* = 62 between variables and emotion language comprehension; 78 between all other variables. All variable scores are raw (unadjusted). BPVS = British Vocabulary Picture Scale. $p<.05$*, $p<.01$**, $p<.001$***.

Table Q2 – TD Sample Partial Correlations Between Emotion Recognition Accuracy, Emotion Language Comprehension, and Receptive Vocabulary (BPVS), Controlling for Age

Variable	1	2	3	4	5
1. Accuracy Total	-	-	-	-	-
2. Accuracy Instrumental	-	-	-	-	-
3. Accuracy Singing	-	.29* (4.38)	-	-	-
4. Accuracy Prosody	-	.35** (7.13)	.33** (1.92)	-	-
5. Emotion Comprehension	.35** (5.23)	.28* (1.08)	.21 (0.41)	.30* (1.57)	-
6. BPVS	.16 (0.27)	.26* (0.92)	-.05 (0.13)	.12 (0.19)	.48*** (>100)

Note. Pearson's correlations. *df* = 59 between variables and emotion language comprehension; 60 between all other variables. All variable scores are raw (unadjusted). BPVS = British Vocabulary Picture Scale. $p<.05$*, $p<.01$**, $p<.001$***.

Appendix R: Chapter 4 Linear Mixed Models for Effect of Language Variables on Emotion Recognition Accuracy, for Referred and TD Samples

Table R1 - Fixed Effects and Marginal R Squared for Linear Mixed Model on Emotion Recognition Accuracy, for Referred Sample

	DV: Emotion Recognition Accuracy	
	X^2	BF_{10}
Model 1		
Intercept	2.87	
Age	0.25	0.35
Sex	1.64	0.63
Condition	40.63***	>100
BPVS	8.03**	3.79
Model Marginal R^2	.22	
Model 2		
Intercept	4.92*	
Age	0.14	0.30
Sex	0.58	0.35
Condition	40.64***	>100
BPVS	1.22	0.49
Emotion Comp.	8.44**	10.36
Model Marginal R^2	.27	
Model 3		
Intercept	4.92*	
Age	0.14	0.40
Sex	0.58	0.36
Condition	41.68***	>100
BPVS	1.22	0.60
Emotion Comp	8.44**	10.57
Emotion Comp*Condition	5.30	0.58
Model Marginal R^2	.28	

Note. $p<.05$*, $p<.01$**, $p<.001$***. BPVS = British Vocabulary Picture Scale. $N = 64$.

Table R2 - Fixed Effects and Marginal R Squared for Linear Mixed Model on Emotion Recognition Accuracy, for TD sample

	DV: Emotion Recognition Accuracy	
	X^2	BF_{10}
Model 1		
Intercept	4.78*	
Age	8.78**	9.61
Sex	0.89	0.43
Condition	59.98***	>100
BPVS	1.23	0.49
Model Marginal R^2	.31	
Model 2		
Intercept	5.28*	
Age	1.93	0.62
Sex	0.85	0.38
Condition	59.98***	>100
BPVS	0.03	0.27
Emotion Comp.	6.31*	3.93
Model Marginal R^2	.35	
Model 3		
Intercept	5.28*	
Age	1.93	0.76
Sex	0.85	0.38
Condition	61.61***	>100
BPVS	0.03	0.34
Emotion Comp	6.31*	3.90
Emotion Comp*Condition	5.28	0.53
Model Marginal R^2	.36	

Note. $p<.05$*, $p<.01$**, $p<.001$***. BPVS = British Vocabulary Picture Scale. $N = 61$.

Appendix S: Chapter 4 Analyses of Interaction Between Acoustic Feature Levels and Emotion Language Comprehension on Emotion Perceptions - Methods

A subset of acoustic features from Chapter 2 were adopted, based on findings in Chapter 2 and those most-examined within past research (e.g., Gabrielsson & Lindstrom, 2010; Llie & Thompson, 2006; Paquetta et al., 2018 - see Appendix B for more information on feature selection). These were mean loudness, mean pitch, tempo/speech rate, and brightness (spectral centroid).

First, to assess relationships between acoustic features and participants' emotion perceptions, z-scored acoustic feature levels for each stimulus was correlated with the proportion of time each emotion was selected. Then, to assess whether individual differences in emotion language comprehension scores related to differences in the links between acoustic features and emotion perceptions, a series of GLMMs were fit. Models were fit for each emotion/condition combination, and each acoustic feature, with emotion selection (yes/no) as the dependent variable. The key independent variable for each model was the given acoustic feature and its interaction with emotion language comprehension scores. Each model also controlled for age and sex and had random intercepts for participant and stimuli.[16]

The entire set of acoustic features were also entered into further GLMMs for each emotion-condition combination, with emotion selection as the dependent variable and participant as a random intercept. The amount of variance in emotion selections explained by the set of acoustic features, while accounting for variance between participants, is presented for these analyses, to contextualise correlations and interactions with emotion language comprehension.

[16] The prosody loudness model had no random intercept for stimuli due to model singularity.

Appendix T: Chapter 4 Analyses of Interaction Between Acoustic Feature Levels and Emotion Language Comprehension on Emotion Perceptions – Results

As Table T1 indicates, there were relatively consistent correlations between loudness and emotion perceptions for emotions other than happiness. Tempo also displayed correlations with all emotions, within at least one condition. Correlations with pitch and brightness were also apparent for some emotions, but these were more sporadic, and in the case of pitch, sometimes in differing directions for different conditions. While accounting for variance between participants, the set of features was best able to predict emotion selections for anger and calmness, across conditions. For singing stimuli, the features only predicted a small amount of variance in selections of fear, sadness, and happiness.

Table T1 – Whole Sample Correlations Between Emotion Perception and Acoustic Feature Levels, and Predictive Value of Acoustic Features, By Condition.

	Emotion Selection														
	Anger			**Fear**			**Sadness**			**Happiness**			**Calmness**		
	Inst.	**Pro.**	**Sing**	**Inst.**	**Pro**	**Sing**	**Inst.**	**Pro.**	**Sing**	**Inst.**	**Pro.**	**Sing**	**Inst.**	**Pro.**	**Sing**
Loudness	.92***	.62**	.92***	.73***	.43^	.07	-.72***	-.30^	-.48*	-.03	.01	-.03	-.82***	-.83***	-.75***
	(>100)	(79.16)	(>100)	(16.21)	(1.28)	(0.58)	(>100)	(0.96)	(8.32)	(0.48)	(0.62)	(0.54)	(>100)	(>100)	(>100)
Pitch	-.42^	.08	-.65**	-.24	.81***	-.38^	-.04	-.09	.28	.43^	.03	.22	.17	-.72***	.63**
	(3.13)	(1.09)	(3.60)	(0.98)	(>100)	(0.82)	(0.54)	(0.58)	(1.16)	(1.18)	(0.52)	(0.98)	(0.92)	(24.56)	(39.39)
Tempo	.41^	.06	.61**	.21	.66**	.55*	-.70***	-.07	-.32^	.51*	.24	.13	-.55*	-.59**	-.79***
	(17.40)	(0.85)	(40.18)	(0.50)	(16.42)	(5.22)	(>100)	(0.48)	(1.58)	(3.89)	(0.52)	(0.48)	(27.54)	(2.15)	(>100)
Brightness	.01	.73***	.76***	.01	-.10	.32	-.33	.08	-.42^	.27	-.24	.07	.01	-.67**	-.77***
	(0.49)	(6.86)	(32.21)	(0.48)	(0.49)	(1.03)	(0.82)	(0.63)	(4.88)	(2.42)	(1.20)	(0.48)	(0.56)	(30.96)	(>100)
R^2	0.36	0.49	0.42	0.14	0.26	0.08	0.37	0.15	0.07	0.23	0.13	0.04	0.31	0.53	0.38

Note. Spearman's correlations. Coefficient (BF_{10}). Data = z-scores. R^2 = marginal value for acoustic features (not including variance explained by random variance between participants). Data points for correlations = each stimulus (n = 20). Data points for R^2 = each trial. Spearman's correlations. r (BF_{10}). Data = z-scores. p<.05*, p<.01**, p<.001***, p<.10^.

A series of GLMMs were then fit to assess whether the relationship between each acoustic feature and emotion perceptions interacted with participants' emotion language comprehension ability. A significant interaction suggests a change in the relationship between an acoustic feature and the perception of the given emotion, as emotion language comprehension ability increases. No interaction suggests that higher emotion comprehension scores did not relate to the strength of relationship between the given feature and emotion selection.

As Table T2 shows, there were differences in the relationship between some acoustic features and emotion perceptions, based on emotion language comprehension ability[17]. For most emotions, within each condition, increasing emotion language comprehension ability was related to a change in the strength of relationship between at least 1 acoustic feature and emotion perceptions[18]. In most of these cases, higher emotion language comprehension ability strengthened the positive or negative relationship between the given feature and the odds of the selection of that emotion. Cases where this was not the case (higher emotion language comprehension related to a weakening or change in direction of this relationship) are highlighted.

[17] Findings are purely correlational, due to separate models being fit for each feature.
[18] Exceptions were fear and sadness for singing stimuli.

Table T2 – Interactions Between Stimulus Acoustic Features and Emotion Language Comprehension Scores, on the Odds of Emotion Selection.

	Emotion Selection*Emotion Comprehension Interaction														
	Anger			**Fear**			**Sadness**			**Happiness**			**Calmness**		
Feature	**Inst.**	**Pro.**	**Sing**	**Inst.**	**Pro**	**Sing**	**Inst.**	**Pro.**	**Sing**	**Inst.**	**Pro.**	**Sing**	**Inst.**	**Pro.**	**Sing**
Loud.	+***	+***	+***	+*	+*	-	-**	-***	-	-	(+*)	+	-***	-***	-**
	>100	>100	>100	1.25	0.94	0.15	11.11	>100	0.20	0.05	0.14	0.06	>100	>100	16.39
Pitch	-**	-	-***	-	+***	-	-	(-***)	+	+***	(+*)	+***	+	-***	+**
	8.00	0.04	>100	0.03	>100	0.07	0.09	3.22	0.04	>100	0.20	4.35	0.03	>100	12.66
Tempo	-	-	+***	+	+***	+	-**	(-**)	-	+***	+***	-	-***	-***	-***
	0.03	0.05	>100	0.03	>100	0.06	9.71	1.29	0.03	>100	9.09	0.03	47.61	45.45	37.04
Bright.	+	+***	+***	+	-	-	-*	-	-*	+*	-	+	+	-***	-*
	0.03	>100	>100	0.02	0.07	0.04	0.48	0.03	0.45	0.50	0.05	0.04	0.03	>100	2.70

Note. $N = 125$. $p<.05$*; $p<.01$**; $p<.001$***. + = positive interaction with emotion comprehension; - = negative interaction with emotion comprehension. Figures = BF_{10s}. Symbol within parentheses = change in direction of effect. Figures = BF_{10s}.

Appendix U: Chapter 5 Robust Multiple Regression Model for Externalising Difficulties Without Prosody Accuracy

	DV: Externalising		
	β	**SE**	**BF_{10}**
Model 1			
Intercept	45.79***	8.78	
Age	-1.63	1.45	0.76
Sex	2.86	2.55	0.45
Household Income	-1.35*	0.58	3.13
Music Training	-1.63	3.05	0.42
Instrumental Accuracy	-16.88*	7.24	3.02
Model Adjusted R^2	.19		

Note. β = unstandardised regression coefficient. $p<.05$*, $p<.01$**, $p<.001$***. $N = 66$.