University of Glasgow | School of Psychology & Neuroscience

VOCAL PERSONALITY AND EMOTION PERCEPTION ACROSS THE EARLY LIFESPAN

Gaby Mahrholz
MA, MSc, MSc(Res)

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

School of Psychology and Neuroscience
University of Glasgow
62 Hillhead Street
Glasgow
G12 8QB

May 2022

University of Glasgow

Declaration of Originality Form

This form **must** be completed and signed and submitted with all assignments.

Please complete the information below (using BLOCK CAPITALS).

Name **GABY MAHRHOLZ**

Student Number

Course Name **PHD PSYCHOLOGY – DOCTORAL THESIS**
Assignment Number/Name **VOCAL PERSONALITY AND EMOTION PERCEPTION ACROSS THE EARLY LIFESPAN**

A link to the University's Statement on Plagiarism is provided at the end of this form. Please read the Statement on Plagiarism carefully THEN read and sign the declaration below.

I confirm that this assignment is my own work and that I have:

Read and understood the guidance on plagiarism in the Student Handbook, including the University of Glasgow Statement on Plagiarism	☒
Clearly referenced, in both the text and the bibliography or references, **all sources** used in the work	☒
Fully referenced (including page numbers) and used inverted commas for **all text quoted** from books, journals, web etc. (Please check with your School which referencing style is to be used)	☒
Provided the sources for all tables, figures, data etc. that are not my own work	☒
Not made use of the work of any other student(s) past or present without acknowledgement. This includes any of my own work, that has been previously, or concurrently, submitted for assessment, either at this or any other educational institution, including school (see overleaf at 31.2)	☒
Not sought or used the services of any professional agencies to produce this work	☒
In addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations	☒

DECLARATION:

I am aware of and understand the University's policy on plagiarism and I certify that this assignment is my own work**,** except where indicated by referencing, and that I have followed the good academic practices noted above

Signed .....

Abstract

The voice is a rich source of social information, and humans have been shown to make quick judgements about trustworthiness and affect perceptions. These initial impressions are proposed to influence our subsequent behaviours and actions towards a person, i.e. whether to approach or avoid them. So far, the literature has investigated vocal trustworthiness perceptions in adult populations, but research exploring its early developmental trajectories is currently missing. Contrastingly, the early maturation of perceived vocal emotion has been studied with a variety of different stimulus types. Yet, there is a gap in the literature for utilising socially-relevant stimuli that are of high ecological validity. Furthermore, listener age has been used as a categorical variable in the majority of research in the field with no consistent age group allocations. This PhD aims to target the gaps in our current understanding of the early developmental trajectories of vocal trustworthiness and affect using age predominantly as a continuous variable. Additionally, we aim to support future research by creating an open-access database of vocal stimuli that are validated on a variety of emotion and personality measures.

Chapter 2 addresses the early development of perceived vocal trustworthiness using emotionally-neutral recordings of the socially-relevant word “hello”. The findings suggest that perceptions of trustworthiness already exist at 5 years of age, however ratings become slightly more positive and “nuanced” with increasing age into early adulthood. In Chapter 3, we expand on the existing literature by investigating developmental patterns of perceived vocal emotion between childhood and early adulthood, using affect representations of the word stimulus “hello”. Findings from this large-scale study suggest that children are able to recognise vocal emotion at higher than chance levels, however, that ability improved significantly with increasing age. We also find that different emotion categories mature at different rates and results are not dependent on either listener sex or speaker sex per se. In Chapter 4, we create the Glasgow vocal emotion and personality corpus starting with affect recordings of the socially-relevant word “hello”. The database is validated on a variety of social measures such as trustworthiness, dominance, attractiveness, affect recognition, recognisability, authenticity, valance and arousal, as well as perceived intensity.

It is openly accessible (with a CC BY 4.0 license), free of charge, and can be used in a variety of settings. Therefore, this corpus is not only a valuable contribution to open science, it also enhances the field by building the foundation for future research aiming to study vocal personality and vocal emotion in unison. Finally, Chapter 5 discusses the implications of the key findings from in this thesis, and highlights limitations and potential future directions for the field.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to thank my supervisor Dr Phil McAleer for his guidance and support on this (extremely) long PhD journey. You were there for me every step of the way, but especially during the pandemic it really felt like you had my back. A huge thanks goes to Dr Heather Cleland Woods who stepped in as a second supervisor during the write-up phase, and for her invaluable input on draft versions and helping me to see "the bigger picture".

I would also like to thank the Glasgow Science Centre for allowing us to hold workshops on emotion perception and collect data. But also a big thanks to the people who didn't have to but still spent hours with me collecting that data: Wilhelmiina Toivo, Greta Todorova, Holly Sneddon, and Greig Dickson. Without you, Chapter 3 would not have been possible (or it would have been extremely thin).

Finally, I want to thank my mum, who has always been encouraging and supportive in everything I do, and friends – old and new. Catrin, my sunshine, for always being there and picking up phone when the house is on fire. And of course, a special thanks to the ~~girls~~ Drs from the Lableless group who I've had the pleasure sharing this journey with. Thanks so much for dragging me across the finishing line this past year. You have been there for the highs and the lows, and I'm immensely grateful to call you my friends.

Last but not least, I want to thank "the best thing that ever happened to me" aka David. You are still my rock, and I could not have done this without your love, support, and patience.

Abbreviations

ACC	Recognition accuracy
AFC	Alternative-forced choice paradigm
CC BY 4.0	Creative Commons Attribution 4.0 International License
CCR	Chance-corrected recognition
CI	Confidence interval
Decoders	Listeners
Encoders	Speakers
GSC	Glasgow Science Centre
H_u	Unbiased hitrates according to Wagner (1993)
ICC	Intraclass correlation coefficient
IEEE	Institute of Electrical and Electronics Engineers
JESS	The Jena Speaker Set
PANGEA	Power ANalysis for GEneral Anova designs
UN	United Nations
VAS	Visual Analogue Scale
WHO	World Health Organisation

Previous Publications

The following chapters are presented as manuscripts that are currently in preparation for publication:

Chapter 2

Mahrholz, G., Greenwood, H., & McAleer, P. (in preparation). Development of vocal trustworthiness perception across the early lifespan

Chapter 3

Mahrholz, G., & McAleer, P. (in preparation). Development of vocal emotion perception across the early lifespan

Chapter 4

Mahrholz, G., & McAleer, P. (in preparation). The Glasgow vocal emotion and personality corpus

Stimuli and validation data have been made available online under a CC-By 4.0 licence (https://osf.io/6da4r/).

In addition, some of the views presented in the general introduction (Chapter 1) have been published in a *research review and opinion piece* in collaboration with Mandy Norrbo and Phil McAleer:

Mahrholz, G., McAleer, P., & Norrbo, M. (2020). Developing voice perception: An overview of current research and models in affect and identity recognition in children and adults. *Cognitive Psychology Bulletin* (5, Spring 2020), 58-65. https://doi.org/10.31234/osf.io/eq2yx

Chapter 1 General introduction

This chapter is structured into the following broad categories: First, the reader is introduced to a general overview of first impressions about speakers, and the current understanding of the developmental trajectories of both perceived vocal trustworthiness (1.1.1) and perceived vocal emotion (1.1.2). Subsequently, section 1.2 highlights challenges and remaining questions within the field and states briefly how this thesis will be addressing them. Finally, section 1.3 summarises the overarching aims of this thesis and emphasizes the contribution and purpose of the individual experimental chapters.

1.1 First impressions about speakers

The voice is a rich source of social information. Humans have been shown to form initial judgements about an unfamiliar speaker after listening to their voice briefly. Such impressions include the speaker's identity (Lavan, Knight, & McGettigan, 2019), gender (Schvartz & Chatterjee, 2012), race (Baugh, 2000; Kushins, 2014), age (Hughes & Rhodes, 2010; Moyse et al., 2014), physical attributes like height and weight (Pisanski et al., 2014; Pisanski et al., 2016; Sell et al., 2010), and emotional states (Banse & Scherer, 1996; Castro & Lima, 2010; Pell & Skorup, 2008). Furthermore, estimating a speaker's intelligence (Schroeder & Epley, 2015), confidence levels (Jiang & Pell, 2015), and personality (Borkowska & Pawlowski, 2011; McAleer et al., 2014) have been shown to be possible from voice-only clues. These impressions are formed even after brief exposure to a speaker's voice. Approximately 500 ms is sufficient to form perceptions about a speaker's personality (McAleer et al., 2014; Mileva & Lavan, 2022) or recognise an expressed emotion with high accuracy (ca. 90%, Lima et al., 2019).

One reason as to why we form initial impressions, in particular trustworthiness, so quickly is to make estimations as to whether to approach or avoid a person (Corr & Krupić, 2017; Engell et al., 2007; Todorov, 2008). From an evolutionary perspective, quickly deciding whether a person should be approached or avoided is of benefit for our immediate survival or self-preservation (Lyon, 2011). Evidence comes from face research when traits and emotion expressions that are related to immediate survival are judged faster than those of no immediate use.

For example, trustworthiness has been reliably judged at less than 100 ms whereas intelligence has not (Bar et al., 2006; Willis & Todorov, 2006). Likewise, emotions have been shown to function as automatic warning systems to facilitate assessments of approach/avoidance (Kamiloğlu et al., 2020; Litt et al., 2020; Mennella et al., 2020; Scherer, 2009). For example, anger and fear alert to threats but are related to different risk assessment strategies (Fessler et al., 2004; Lerner & Keltner, 2001). Experiencing anger signals the assessed risk as low and therefore motivates confrontational approaching behaviour, whereas for fear, risk is assessed as high and promotes withdrawal (Darwin, 1872; Moons et al., 2010; Shariff & Tracy, 2011). Similarly, experiencing disgust warns of aversive foods or distasteful ideas and behaviours and triggers avoiding behaviours, such as pathogen avoidance (Cepon-Robins et al., 2021; Darwin, 1872; Shariff & Tracy, 2011).

A slightly different framework was proposed with the emotion overgeneralisation hypothesis (Todorov, 2008). According to this model, stable personality traits are difficult to assess from brief exposure, therefore making judgements about someone's personality hinges upon momentary, dynamic emotion perceptions. We subsequently rely on those generalisations to estimate whether someone is trustworthy enough to be approached or best be avoided. Regardless whether trustworthiness is assessed separate from emotion or in combination, the emphasis lies on the approach/ avoidance theorem to guide our subsequent behaviours and actions towards a person.

To be able to extract information quickly and detect further intentions, we have to rely not only on what is said (i.e. linguistic features of speech) but how something is said (i.e. interpreting para- and extralinguistic cues) since much meaning is conveyed through non-verbal parts of language (Burgoon et al., 2010). Laver (1994) describes paralinguistic cues as non-linguistic. Similar to linguistic features, they are coded information in which the speaker is aiming to get a certain message across, by placing emphasis or expressing emotions, arousal, and attitudes (Kreiman & Sidtis, 2011; Laver, 1994; Schweinberger et al., 2014). Beyond the auditory dimension, paralinguistic cues may include gestures, posture, body-movement, facial expressions, etc. (Laver, 1994). Finally, Laver defines extralinguistic cues as "residue of the speech signal" (p.

22) after linguistic and paralinguistic information has been accounted for. He describes these cues as the non-coded part of language that provides information about a speakers' identity, such as speaker's voice quality, and overall range of pitch and loudness. Schweinberger et al. (2014) slightly expands on Laver's (1994) definition by relating extralinguistic cues to the more stable speaker characteristics (e.g. identity, biological sex, social gender, age, and socioeconomic background), and link paralinguistic cues to dynamic information (e.g. emotion).

However, Laver (1994) also states that extracting information about a speaker might not map exactly onto the three types of speech distinctively. Whilst listeners may extract information about emotional states predominantly from paralinguistic cues, perceived personality may be extracted from "any or all" (Laver, 1994, p. 23) types, which leaves room for errors in initial judgement. In fact, it has been debated whether these initial impressions accurately relate to the true representation of a speaker (Klofstad & Anderson, 2018; Zebrowitz & Collins, 1997; Zebrowitz & Montepare, 2008). Here we argue that this debate is secondary, since our behaviours and actions towards a speaker are guided by the first impression we form about them and not their actual personality or emotional state. Regardless, these initial impressions have been shown to be highly consistent between listeners (Baus et al., 2019; Mahrholz et al., 2018; McAleer et al., 2014; Oleszkiewicz et al., 2017; Schirmer et al., 2019), and stable within listeners between brief and prolonged exposure as well as different speech contents (Mahrholz et al., 2018).

1.1.1 Perceived vocal trustworthiness and its early developmental trajectories

Trust is one of the fundamental building blocks to develop and maintain happy, well-functioning relationships (Simpson, 2007). In voice research, trustworthiness has been identified as one of the two key dimensions in the social voice space model (McAleer et al., 2014). As mentioned earlier, listeners show remarkable agreement between each other and consistency in evaluating who sounds trustworthy and to which degree (Baus et al., 2019; Mahrholz et al., 2018; McAleer et al., 2014). Furthermore, trustworthiness judgements are made similarly between blind and sighted people (Oleszkiewicz et al., 2017). This

could suggests that first impressions are based on an internal protype, similar to what has been suggested for identity (Latinus & Belin, 2011a).

How vocal trustworthiness develops across the early lifespan has not received much attention in research. We are only aware of one paper (Schirmer et al., 2019)[1] that investigated the relationship between vocal trustworthiness and age, however they opted for a comparison between younger and older adult groups. The study reported main effects for speaker sex and speaker age: younger speakers and female speakers were perceived as more trustworthy than older or male speakers respectively. Listener age or listener sex effects were non-significant. Yet, how the percept of vocal trustworthiness develops and matures between childhood and adulthood remains unknown.

Turning to other vocal traits for a comparison, it appears that developmental aspects utilising vocal stimuli have not been thoroughly investigated either; bar attractiveness. In a series of studies and experiments, Saxton and colleagues (2006; 2009; 2013) found that lower-pitched male voices were rated as significantly more attractive by adolescent female listeners who have entered puberty and by adults. Younger girls who had not entered puberty yet, showed no preference towards lower- or higher-pitched male voices. For male listeners, however, the results were a bit more ambiguous (Saxton et al., 2009, 2013). Younger boys rated higher-pitched female voices as more attractive compared to lower-pitched versions of the same voice. Contrastingly, older boys showed no particular preference for either higher- or lower-pitched female voices. This is in contrast to adult men, who frequently judge a higher-pitched female voice as more attractive (Borkowska & Pawlowski, 2011; Collins & Missing, 2003; Feinberg et al., 2008; Jones et al., 2008). Saxton et al. (2009, 2013) offered the explanation that adolescent males could be more attracted to the lower-pitched female voice since pitch drops for females during puberty. This might subsequently be taken as an indication of sexual maturation by the older boys. Though, this does not entirely explain how, why, and when preferences shift back to higher-pitched female voices. Regardless of the ambiguity in internal mechanisms within adolescent males, what could be reasoned from Saxton's

[1] Now retracted due to errors in the acoustic analysis. Here we still acknowledge and value the contribution from the behavioural analysis.

research is that perceived attractiveness becomes more important during puberty, i.e. when mate-selection becomes relevant.

However, when presenting the British stimuli to Czech participants, results could not be replicated for either the girls or the boys (Saxton et al., 2010). This may hint at cross-cultural differences, and suggests that other influences such as society or environment may shape who we might find attractive or not. Conversely, a cross-cultural difference does not appear to be mirrored in vocal trustworthiness perception. For example, Baus et al. (2019) had Spanish and UK participants rate native and foreign language respectively, and found that trustworthiness of a speaker was perceived similarly regardless of culture. Furthermore, vocal attractiveness of Spanish female and male speakers mapped onto Component 1 (related to trustworthiness) in a Principal Component Analysis (PCA), which is similar to findings from Scottish female speakers (McAleer et al., 2014). Contrastingly, for Scottish male speakers, perceived attractiveness was mapped onto Component 2 (related to dominance). However, the studies by Baus et al. (2019) and McAleer et al. (2014) were conducted in adult listeners, therefore it is difficult to draw developmental conclusions. Yet, taken together with findings by Saxton et al. (2010), it challenges how similar the vocal traits of trustworthiness and attractiveness are to draw meaningful associations.

To obtain further insights into the developmental trajectories of vocal trustworthiness, considering evidence from face research may be beneficial since voice and face perception share some similarities (McAleer et al., 2014; Young et al., 2020; but Lavan, Burton, et al., 2019). Trustworthiness perception from emotional neutral faces appears to develop early in life. Six- to 8-month old infants have shown a preference for trustworthy-looking faces over untrustworthy-looking ones (Jessen & Grossmann, 2016; Sakuta et al., 2018). However, Sakuta et al. (2018) showed that the preference for computer-generated trustworthy faces appeared only when stimuli were also of high dominance. For faces low in dominance, there was no preference of trustworthy over untrustworthy-looking faces. Jessen and Grossmann (2016) did not identify a preference for dominant vs non-dominant faces in itself, implying a complex interaction between trustworthiness and dominance in infants. Overall, this suggests some sort of innate ability for trait perceptions is present in early life,

however this ability is unlikely of adult-like consensus given that trustworthiness and dominance have been proposed as independent dimensions for adult perceptions in face (e.g. Todorov et al., 2008) and voice research (e.g. Baus et al., 2019; McAleer et al., 2014).

At around 3 to 4 years of age, children seem able to make explicit nice/mean judgements of computer-generated faces at better than chance-level, however adult-like consensus was only shown for 5- to 6-, and 7- to 10-year-olds (Cogsdill et al., 2014). Similar results have been reported when judging how nice/mean natural face stimuli looked (Cogsdill & Banaji, 2015). Adults scored significantly higher accuracy for adult and macaques faces (but not for children's faces) compared to 3- to 13-year-olds (mean age = 6 years). However, the children's age range of 3 to 13 years is exceptionally large to detect potential subtle developmental patterns. Given the mean age of 6 years, the results may have been biased towards younger children's perception ability.

When using more complex rating paradigms requiring higher cognitive load (i.e. explicit trustworthiness rating task with a 9-point Likert scale), Caulfield et al. (2016) found that participants of 5, 7, and 10 years of age, and adults were able to reliably distinguish between trustworthy and untrustworthy faces at better-than-chance level. Yet, a rating pattern emerged for the 5- and 7-year-olds that was significantly different to the 10-year-olds and adults. The younger two age groups rated the trustworthy-looking faces as lower, and the untrustworthy-looking faces as higher in trust compared to the older children and adults (i.e. bias towards the mean). The 5- and 7-year-olds did not differ significantly in their responses; neither did the 10-year-olds and adults, suggesting ongoing maturation between 7 and 10 years of age.

Further evidence from economic game paradigms (Charlesworth et al., 2019; Ewing et al., 2015; Ewing et al., 2019) indicates that children start acting upon the information they perceive around 5 years of age. They begin tailoring their subsequent behaviours towards people, for example by presenting gifts towards the more trustworthy faces (Charlesworth et al., 2019), or making associations to other traits. Palmquist et al. (2020) showed that puppets with highly trustworthy-looking headshots of natural faces were perceived as more knowledgeable and competent than their untrustworthy-looking counterparts.

Overall, these findings seem to suggest that there are some innate components to facial trustworthiness. Yet, the ability to make explicit judgements at better-than-chance levels, appears at around 3 to 4 years of age when rating tasks are employed with low cognitive load (i.e. a 2-AFC) or between 5 and 7 years of age with slightly more complex rating paradigms (i.e. Likert scale responses, economic games). These abilities seem to fine-tune throughout the early life-span until adult-like consensus is reached at around 10 years of age. However, recently, researchers (Lavan, Burton, et al., 2019; Young et al., 2020) have started to emphasize the dissimilarities between the vocal and facial domains, highlighting modality-specific characteristics (e.g. voice stimuli are always dynamic whereas face stimuli could be either static or dynamic) and physical properties of the stimuli (e.g. fundamental frequency vs visual contrasts). Since vision and audition are two separate modalities, they may not be assumed to have exactly the same processing stages between them. This may subsequently impact on perception and its developmental trajectories. Therefore, whether perceived vocal trustworthiness mirrors the developmental trajectories of facial trustworthiness exactly remains to be investigated.

1.1.2 Perceived vocal emotion and its early developmental trajectories

In his seminal work, Darwin (1872) was the first person to note facial emotion categories and assuming a biological basis for them. Darwin's insights stemmed predominantly from observing children and adults, and noticing similarities in facial expressions generated. Despite being incorrectly critiqued for his scientific methods at times (see Ayala, 2009, for a review), Darwin's observations lay the groundwork for others to build upon. For example, Ekman and Friesen (1971) created the "Basic Emotion Theory" emphasising the innate psychological and biological components of facial expressions. Their influential theoretical framework proposes there are six basic overarching "emotion families" – namely happiness, sadness, anger, surprise, fear, and disgust. These discrete emotion families are suggested to be distinct and distinguishable from one another and are universally recognised. Whilst the universality aspect of facial emotion has recently been questioned (e.g. Chen & Jack, 2017; Jack et al., 2012; Jack et al., 2016), these six basic emotion categories have been shown to replicate with vocal stimuli and across different cultures/ languages (Bryant & Barrett, 2008;

Chronaki et al., 2018; Kawahara et al., 2017, 2021; Laukka & Elfenbein, 2021; Pell & Skorup, 2008; Sauter, Eisner, Calder, & Scott, 2010; Sauter, Eisner, Ekman, & Scott, 2010; Scherer et al., 2001).

The development and maturation of vocal emotion across the early lifespan has received more attention than vocal trustworthiness. Research has suggested that rudimentary emotion detection skills from voices emerge gradually during infancy and toddlerhood (e.g. Flom & Bahrick, 2007) which may be similar to the trustworthiness perceptions outlined earlier. However, the question remains when vocal emotion recognition starts to differ from chance levels, and when adult-like recognition rates are being achieved.

Toddlers (2- to 3-year-olds) and pre-schoolers (3- to 5-year-olds) seem not very accurate at recognising emotion in voice-only scenarios, despite being able to identify face-only, body-only or multi-cue stimuli with high accuracy (Chronaki et al., 2015; Nelson & Russell, 2011; Quam & Swingley, 2012). This has been suggested by research using many different designs and stimulus types. For example, Chronaki et al. (2015) used a morphing paradigm (i.e. morphs of angry, happy, sad with neutral at 50%, 75%, and 100% morphs) for non-verbal interjection (i.e. ‘ah’) and found discrimination accuracy for 3.5 to 5.5 year-old pre-schoolers was around chance levels. Whilst the numbers reported in the paper improved slightly with increasing intensity, the authors did not report whether this improvement was significantly different from chance. Comparable results were obtained by Aguert et al. (2013) when presenting a semantically-neutral cartoon to 5-year-olds using unintelligible speech. Pre-schoolers’ performance to estimate whether character “Pilou” felt good or bad did not differ significantly from chance. The authors speculated that 5-year-old children may take additional information from a seemingly neutral scene into account rather than focussing solely on the prosody of the speaker. However, it may be considered that the designs by Chronaki et al. (2015) and Aguert et al. (2013) are quite complex and potentially require higher cognitive load which could have contributed to the low recognition rates in pre-schoolers.

Research with less complex paradigms suggests that accuracy for pre-schoolers around 5 years of age is better than chance levels. Quam and Swingley (2012) used non-verbal vocalisations (‘mmm, mm mm mmm’; Experiment 1) or

semantically-neutral sentences ('Oh look at that'; Experiments 2 and 3) in a puppet show set-up that were either produced live (Experiments 1 and 2) or pre-recorded (Experiment 3). Children then voted in an alternative forced choice (AFC) paradigm whether "Puppy" was happy or sad to have found the toy of choice or a different toy respectively. Overall, the study found 3-year-olds' recognition accuracy was either at or below chance level for the non-verbal and sentence condition, improved slightly for the 4-year-olds, and increased further for the 5-year-olds. Similar findings were reported by Nelson and Russell (2011) and Zupan (2015) also using semantically-neutral sentences (Nelson & Russel: "I felt this feeling before; it was just a few days ago"; Zupan: "I'm going out of the room now and I'll be back later."), whilst including 4 emotion categories of happiness, sadness, anger, and fear. Despite using different paradigms (Nelson & Russel: free-response; Zupan: 4-AFC), both studies found that pre-schoolers recognised vocal emotion accurately above chance levels. Zupan (2015) found results held true for different levels of intensity, reporting 5-year-old pre-schoolers' recognition accuracy at 49% for low- and at 61% for high-intensity vocal stimuli. Taken together these findings seem to suggest that vocal emotion recognition may start to differ significantly from chance at around 5 years of age, though adult-like recognition rates are not yet reached. However, results are fairly ambiguous as to whether subtle emotions can be detected accurately at age 5. Perhaps pre-schoolers are able to detect low-intensity naturally produced emotions (e.g. Zupan, 2015) but may struggle with more complex morphing paradigms (e.g. Chronaki et al., 2015).

Beyond age 5, a large body of research suggests that recognition rates increase steadily (e.g. Allgood & Heaton, 2015; Chronaki et al., 2015; Doherty et al., 1999; Sauter et al., 2013; Zupan, 2015). Using an alternative forced choice paradigm, Sauter et al. (2013) showed that children between 5 and 7 years of age recognised emotions in non-verbal stimuli (laughs, cries, grunts; 78% accuracy) and inflected speech (three-digit numbers; 53% accuracy). This increased to 84% and 72% respectively for the 8- to 10-year-old age group. The study included 10 emotion categories but only presented 4 options at the time to limit cognitive load (i.e. chance levels of 25%). When using a 3-AFC task with emotion categories happy, sad, and fearful (i.e. chance levels of 33%) of the same non-verbal stimuli by Sauter et al. (2013), Allgood and Heaton (2015) found

52% recognition accuracy in 5- to 6-year-olds which increased to 63% (age group 7-8 years) and 80% (age group 9-10 years). However, the recognition rates in Allgood and Heaton (2015) may have been inflated for the 5- to 6-year-olds, since approximate 30% of participants were excluded before data analysis due to performing at chance levels.

A slightly less steep improvement during childhood was suggested in a study by Doherty et al. (1999) using a 3-AFC design including emotion categories happiness, anger, and sadness. Recognition accuracy for the 5.5-, 6.5-, 7.5-, and 8.5-year-olds was reported at 66.2%, 80.8%, 82.5%, and 93.3% respectively. Contrastingly, the study by Tonks et al. (2007) found no significant differences in recognition accuracy between the age groups of 9, 10, 11, 12, 13, and 14+ years in a 5-AFC paradigm with emotion categories happiness, anger, fear, sadness, and neutral. The average recognition accuracy across all 6 age groups was 84% (SD = 12.9%; min = 78% at 9 years; max = 88% at 11 and 12 years). However, the non-significant results could be due to the study's low sample size (ca 10 participants per age group) and being most likely underpowered to detect very small effect sizes. Additionally, neither Doherty et al. (1999) and Tonks et al. (2007) included an adult comparison group. Whilst the findings suggest that emotion accuracy is high at the end of childhood, the question remains whether it is at adult-like ability or how accuracy develops during adolescence.

When adult comparison groups are included, research has suggested that vocal emotion recognition continues to mature into adolescence or even beyond, though there is ambiguity in the literature as to when adult-like consensus is reached. Some research (Brosgole & Weisman, 1995; Zupan, 2015) found adult-like accuracy is reached by early-adolescence (around 12 years of age). However, this is in slight contrast to research still finding significant differences between older children (8.5 to 10.5 years; Chronaki et al., 2015)/ adolescents (11 to 13 years, Chronaki et al., 2018; or 13 to 15 years, Morningstar, Ly, et al., 2018) and adult comparison groups. This is further corroborated by Aguert et al. (2013) reporting significant differences in vocal emotion accuracy between 5- and 9-year-olds, and 9-year-olds and adults, but non-significant differences between 9- and 13-year-olds. The study used the "9 year" age group as baseline, thus differences between the 13-year-olds and adults were not commented on.

However, looking at the recognition accuracy for 5-, 9-, 13-year-olds, and adults at 51.7%, 79.2%, 80.8%, and 96.7% respectively could suggest continuous development into young adulthood. Further evidence for ongoing maturation beyond early adolescence is suggested by Grosbras et al. (2018). The study found that emotion recognition follows a quadratic pattern with steeper improvement between childhood and adolescence, and adult-like performance levels being reached between 14 and 15 years of age. In support of maturation further into adulthood, Amorim et al. (2021) showed further ongoing maturation between 15 and 23 year-old participants. This suggests that development may continue into late adolescence or early adulthood but perhaps with more noticeable growth patterns during early adolescence.

From this encounter it is difficult to predict when actual developmental milestones are reached. Perhaps something is actively happening at 5 years of age but with easier paradigms and better distribution of cognitive load, younger children may be better than chance. However, whether emotion recognition abilities are adult-like by early, mid, or late adolescence, or whether development continues into early adulthood, may depend on a variety of factors, and remains to be investigated. Those findings imply that vocal emotion development continues to mature between late childhood/ early adolescence and adulthood.

1.2 Challenges and remaining questions

1.2.1 Gap in the literature for early development of perceived vocal trustworthiness

As outlined previously, no research has investigated how perceived vocal trustworthiness develops across the early lifespan. To establish a rationale, developmental patterns were reviewed in vocal attractiveness and facial trustworthiness research.

To briefly summarise, the development of vocal attractiveness seemed to reach a key milestone when mate selection becomes relevant (e.g. Saxton et al., 2010). However, it was questioned whether this milestone is mirrored for vocal trustworthiness when research suggests significant cross-cultural differences of

vocal attractiveness perceptions (Saxton et al., 2010) but not of perceived vocal trustworthiness (Baus et al., 2019; McAleer et al., 2014), albeit in different populations (i.e. adolescents vs adults). Furthermore, McAleer et al. (2014) showed that vocal attractiveness mapped predominantly onto Component 1 (related to trustworthiness) for Scottish female speakers but onto Component 2 (related to dominance) for Scottish male speakers in the social voice space model. The study also computed step-wise regression and found that the combination of components 1 and 2 explained 54% and 66% of the variance in attractiveness for male and female speakers respectively. Overall this seems to suggest that vocal attractiveness and trustworthiness appear not too closely related with one another in adult populations, hence questioning whether they share developmental trajectories.

Analogies were subsequently sought in judgements of facial trustworthiness given the similarities between the vocal and facial dimensions usually emphasized (McAleer et al., 2014; Young et al., 2020). Facial trustworthiness perceptions start differing from chance at around 3 to 4 years of age for simple task designs, but 5 to 7 years for when paradigms require higher cognitive load. These abilities seem to further fine-tune throughout the early life-span until adult-like consensus is reached at around 10 years of age.

Yet, recently the focus has shifted from highlighting the commonalities to emphasizing the dissimilarities between the vocal and facial domains (Lavan, Burton, et al., 2019; Young et al., 2020). These differences may question whether perceived facial trustworthiness and its developmental trajectories are mirrored exactly on the vocal domain. This thesis therefore sets out to close the gap in the developmental literature of perceived vocal trustworthiness.

1.2.2 The need for socially-relevant word stimuli and an open-access database

Social traits so far have been studied using vowel sounds, multiple-vowel series, words, sentences, and longer passages (e.g. Baus et al., 2019; Ferdenzi et al., 2013; Lavan et al., 2021; Lavan et al., 2020; Mahrholz et al., 2018; McAleer et al., 2014). However, it is very typical for researchers in the field of vocal personality perceptions to record stimuli required for a specific experimental

set-up, and as a result not many pre-validated stimuli are openly shared with other researchers. Currently, there are three open-access databases available that include stimuli pre-validated on social traits. One is a collection of vowels and sentences with neutral meaning in French that were validated on trustworthiness, dominance, attractiveness, masculinity/ femininity, beauty, and health (Ferdenzi et al., 2015). The second corpus is a collection of German sentences, syllables, read text, semi-spontaneous speech, and vowels validated on perceived attractiveness, likeability, distinctiveness, regional accent, and age (Zäske et al., 2020). The only database with stimuli in English including words and sentences with socially-relevant and -ambiguous meaning were validated on the dimensions of perceived trustworthiness, dominance, and attractiveness (Mahrholz et al., 2018). All stimuli in those vocal personality corpora are emotionally neutral representations of speech.

When studying vocal emotion perception, researchers have used a variety of stimuli, such as non-verbalised emotional interjections/ affect bursts such as cries, laughs, and grunts (e.g. Allgood & Heaton, 2015; Amorim et al., 2021; Belin et al., 2008; Chronaki et al., 2015; Laukka et al., 2013; Sauter & Scott, 2007), nonsense syllables (e.g. bábaba; Mildner & Koska, 2014), pseudo-words (Demenescu et al., 2014), three-digit numbers (Sauter et al., 2013), pseudo-sentences (Aguert et al., 2013; Banse & Scherer, 1996; Chronaki et al., 2018; Paulmann & Pell, 2010), sentences with neutral content (Kawahara et al., 2021; Morningstar, Ly, et al., 2018; Nelson & Russell, 2011; Zupan, 2015), or sentences employing congruous/ incongruous paradigms (e.g. happy content with happy/angry prosody; Friend, 2000; Morton & Trehub, 2001). Subsequently, many researchers have made their stimuli available (either open-access or upon-request) for other researchers to use (e.g. Belin et al., 2008; Burkhardt et al., 2005; Castro & Lima, 2010; Lassalle et al., 2019; Laukka et al., 2010; Laukka et al., 2013; Lima et al., 2013; Sauter, Eisner, Ekman, & Scott, 2010; Sauter & Scott, 2007; Schirmer et al., 2019). Some multi-modal corpora consist of additional facial stimuli, but again, the vocal stimuli include either sentences or vowel sounds (Bänziger et al., 2009; Bänziger et al., 2012). Socially-relevant word stimuli are currently missing from openly available databases.

It may be that the chosen stimulus type contributes to the ambiguity in the literature as to when developmental milestones are being reached. Emotion recognition from non-verbal vocalisations is often labelled “easier” and recognition rates are higher compared to emotion recognition from prosodic speech (Hawk et al., 2009; Hunter et al., 2010; Laukka et al., 2013; Sauter et al., 2013). Direct comparison between affect bursts and inflected speech implies this holds true in children (Sauter et al., 2013) and in adults (Hawk et al., 2009). When employing complex paradigms or nuanced expressions, i.e. morphing to decrease emotional intensity or introducing incongruency, children tend to have lower recognition rates than other age groups (Chronaki et al., 2015; Friend, 2000; Morton & Trehub, 2001). Morningstar, Nelson, and Dirks (2018) have suggested that nuanced expressions may only be mastered with additional maturation or further proficiency to accurately interpret them.

Furthermore, the ecological validity of currently utilised stimulus types may be questioned. Affect bursts and longer speech scenarios, in particular with incongruent paradigms, are very common in everyday speech, and therefore high in ecological validity. Other stimulus types, such as three-digit numbers or syllables, are perhaps less applicable to daily social encounters. Furthermore, non-verbal affect bursts could still be of considerate duration (~1000 ms; e.g. Amorim et al., 2021; Belin et al., 2008; Sauter et al., 2013), and be overly theatrical (for example the MAV; Belin et al., 2008) which, again, could impact ecological validity. Thus, it is important for the field to use brief speech stimuli that are socially meaningful to the listener because they are encountered frequently in daily life. Whilst the socially-relevant word stimulus “hello” is already used in vocal personality perception research with adult listeners, none of the existing openly-available emotion corpora include brief speech stimuli that are of social relevance. Hence, we argue that vocal emotion research would also benefit from this speech stimulus with high ecologic validity.

In this thesis, we will therefore focus on brief representations of the socially-relevant semantically-neutral word “hello” to study the developmental trajectories of perceived vocal trustworthiness and emotion. Stimuli will be varying in intensity to increase ecological validity. Once validated on a variety of

personality and emotion dimensions, these stimuli will be published in an open-access database of affect vocalisations.

1.2.3 Gender differences in listeners and speakers

When studying vocal trait perceptions in adult populations, commonly no main effect of listener sex is observed (e.g. Amir & Levine-Yundof, 2013; Baus et al., 2019; Bruckert et al., 2010; Mahrholz et al., 2018; McAleer et al., 2014; Schirmer et al., 2019; Zäske et al., 2020). However, one study found male listeners rated male speakers as less attractive than female listeners did (Babel et al., 2014). Nevertheless, the same overall pattern emerged for male and female listeners as to which speakers sounded more/less attractive. The authors interpreted this finding as an unwillingness by male listeners to rate male speakers on an attractiveness dimension rather than a distinct gender difference in perception. Despite not detecting a main effect of listener sex, Zäske et al. (2020) reported interactions of listener sex and speaker sex with speaker age (younger vs older adult speakers) for attractiveness and likability. However, it is difficult to speculate whether or how these findings translate to the early developmental trajectories of perceived vocal trustworthiness.

In contrast to personality research, there is a lot of ambiguity in the literature surrounding gender differences of either listeners or speakers when investigating vocal emotion perceptions. Some research has reported an overall advantage of female listeners decoding information (Belin et al., 2008; Collignon et al., 2010; Grosbras et al., 2018; Keshtiari & Kuhlmann, 2016; Paulmann & Uskul, 2014; Sen et al., 2018), however, other findings suggest that these listener sex differences are either small (Lausen & Schacht, 2018; Thompson & Voyer, 2014) or non-significant (e.g. Amorim et al., 2021; Lima et al., 2014; Paulmann et al., 2008; Sauter et al., 2013). It may be suggested that some particular emotion categories drive the effects of listener sex. For example, Sen et al. (2018) showed a female advantage for happy and neutral categories but females' and males' accuracy was on par for anger and fear representations.

Turning to speaker sex differences, the results are equally ambiguous. Some describe that females are the better encoders of emotion in affect bursts (Belin et al., 2008; Lausen & Schacht, 2018) and pseudo-words (Lausen & Schacht,

2018), however, males seem to portray negative nouns better than females (Lausen & Schacht, 2018). For other stimuli types, such as semantically positive or neutral nouns, Lausen and Schacht (2018) reported non-significant differences. Factors that could explain the ambiguity are not readily detectable. Results from Lausen and Schacht (2018) may suggest stimulus choice plays a role, but future research needs to investigate further with explicit paradigms.

It needs mentioning that gender is not the main variable of interest within this thesis. However, given gender could potentially influence findings, listener sex, speaker sex, and interactions with other variables will be included within all statistical models throughout this thesis.

1.2.4 Different maturation rates for different emotion categories

Another reason that could play a role as to when adult-like consensus is reached could be the inclusion or exclusion of certain emotion categories. A closer look at Morningstar, Ly, et al. (2018) showed that the age effect between mid-adolescents was driven by significant differences of fear and sadness. No significant differences were found for anger, happiness, or disgust. This in contrast to Nelson and Russell (2011) who showed pre-schoolers were better at recognising sadness than anger, happiness, and fear. Yet, whilst all emotion categories were recognised better by the adult comparison group, sadness and anger seemed to improve less than happiness and fear. Data from Kawahara et al. (2017; voice-only data requested from the authors) suggests a slightly different pattern as the accuracy of Japanese listeners' (age groups: 5-6, 7-9, 10-12, and adults) improved continuously with each age group for anger but not for happiness. Happiness was highly recognised across all age groups (~90%). Whilst Grosbras et al. (2018) agree with Kawahara et al. (2017) that happiness was the most, and anger the least recognised emotion category, their results suggested that all emotion categories were improving with an increase in age. Amorim et al. (2021) included a variety of basic and complex emotion categories and found no significant differences between childhood and adolescence for pleasure, relief, sadness, and surprise. However, accuracy of all other emotion categories (achievement, anger, disgust, fear, happiness, and neutral) improved. Though, significant changes between adolescence and young adulthood seemed to be driven by fear and achievement. This suggests that some emotion

categories may develop faster than others, yet, there is ambiguity over which exact emotion categories contribute to which extend.

This thesis will therefore investigate all six basic emotion categories, rather than focussing on specific subsets, and consider the interaction between age and emotion category.

1.2.5 Operationalising age

The majority of emotion research has investigated developmental aspects by allocating participants into age groups. This section will focus on the challenges that researchers face operationalising age.

1.2.5.1 Terminology and definition of age groups

Depending on textbooks and articles, early childhood starts at around 2 (Lally & Valentine-French, 2019) or 3 years of age (Santrock, 2020) and lasts until age 5 or 6 (Lally & Valentine-French, 2019; Santrock, 2020). The period starting around 5 to 6 years of age and ending approximately between age 10 and 12, is sometimes referred to as middle childhood (Berk, 2017; Rathus, 2017; Shaffer & Kipp, 2014), and sometimes labelled middle to late childhood (Lally & Valentine-French, 2019; Santrock, 2020). The Centers for Disease Control and Prevention (CDC, 2021, February 22), confusingly, distinguish between middle childhood (6-8 years) and middle childhood (9-11 years).

Adolescence is suggested to start from the onset of puberty and end in the late teens/early twenties (Blakemore, 2018). Santrock (2020) provides age ranges from between 10 and 12 years of age to approximately 18 to 21 years. However, the World Health Organisation (WHO, 2022a) opts for more rigid age limits and defines adolescence as the period between 10 and 19 years, though they acknowledge that there may be a developmental difference between a 10- and a 19-year-old. Therefore, they suggest breaking down adolescence into early (10–13 years), middle (14–16 years), and late (17–19 years) stages, yet, in one of their most recent publications, the WHO (2021, November 17) only distinguish between younger (10-14 years) and older adolescents (15-19 years). Similar to the latter categorisation, the CDC (2021, February 22) separates their adolescent age range into “Young Teens” (12-14 years) and “Teenagers” (15-17 years).

Young adults are subsequently defined as persons between 20- to 34 years of age (ONS, 2016, February 22), or more broadly as starting in late teens/early twenties and lasts through the thirties (Santrock, 2020; Shaffer & Kipp, 2014). Sometimes this period is defined to last into the mid-40s (Lally & Valentine-French, 2019). Recently, the age range between adolescence and young adulthood has been the centre of debate. One argument is focused on the delayed shift in social role transitioning, for example from education into the job market, seeking permanent partnerships and considering marriage, committing to parenting, or moving out of the family home (Arnett, 2016a, 2016b; ONS, 2016, February 22; Santrock, 2020; Sawyer et al., 2018). Since the onset of these adult commitments is happening later in life now compared to 50 years ago, they could delay the psychological development necessary to accomplish these steps (Arnett, 2016a). A second argument is built around evidence from the neuroscience literature and the continued neurological maturation of the *social brain network* into the mid-20s (Dosenbach et al., 2010; Kilford et al., 2016). For example, the pre-frontal cortex, involved in decision-making and executive functioning, has been shown to develop until around 25 years of age (e.g. Arain et al., 2013).

Since global life expectancy has increased by 6.5 years within the last 20 years (WHO, 2022c, May 20), and to incorporate the changes in societal and biological development, some authors (Patton et al., 2018; Sawyer et al., 2018) propose the adolescence age range should increase proportionally and include persons between 10 and 24 years of age. Kinghorn et al. (2018) acknowledge the crucial period between 10 and 24 years of age, however they suggest a “disaggregation” into young (10–14 years), middle (15–19 years), and late (20–24 years) adolescence to “support evidence-based interventions and policies” (Kinghorn et al., 2018, p. e10). Whilst the age ranges partially overlap with the latest reports published by the WHO (e.g. 2021, November 17), terminology does not. In general, the WHO (2022a) promotes the terminology of “young people” for 10- to 24-year-olds. Arnett (e.g. 2016a), however, proposes a separate transitioning period between 18 and 25 years to be called “emerging adulthood”, though the Society for the Study of Emerging Adulthood (SSEA, 2014) suggests this transitioning period should include age ranges between 18 and 29 years. Other researchers question the generalisability of the Emerging Adulthood Theory

altogether, as cultural aspects, social classes, or ethnicity may influence the span in these age ranges considerably (du Bois-Reymond, 2016; Furstenberg, 2016; Silva, 2016).

To make matters more complex, other sources (United Nations, n.d.; see also WHO, 2022a) use terminologies and definitions that differ quite substantially from the classifications discussed above. They categorise a person between 15 and 24 years of age as “youth” which results in a “child” defined as being 14 years or younger. According to the United Nations (n.d.), “youth” would therefore include the majority of the group they define as “Teenagers” (13-19 years) and the “Young Adults” (20-24 years). Turning toward legal terminology, the Convention on the Rights of the Child (United Nations, n.d.) defines any person under the age of 18 years as “child” which would imply that the majority of adolescents fall into this category. Since no Convention on the Rights of the Adolescence exists, this was deliberately done by the United Nations to ensure the legal rights of adolescents.

1.2.5.2 Consequences for research

Defining developmental stages according to physiological, cognitive, social, and emotional milestones may have its merits. However, assigning concrete age ranges to these developmental stages accurately proves challenging for researchers in reality. Is a hypothetical 11-year old to be sorted into the child or adolescent age group, and would that be influenced by whether they have reached puberty yet? Attempting to broadly categorise age may result in the loss of detailed information or a decrease in power (e.g. Altman & Royston, 2006). It has also been argued that participants on either side of the grouping borders may be more similar than they are different (e.g. Sauerbrei & Royston, 2010).

Yet, as outlined above, perceived vocal personality and emotion have predominantly been studied by allocating participants into groups according to chronological age. Having to make decisions with too much flexibility in different fields, may have contributed to the substantial variability of developmental findings within the literature. Not only are outcomes of studies and experiments difficult to compare, it may subsequently impact whether group differences are significant or not depending on how particular age ranges

are defined. Turning toward a continuous development approach would be more beneficial for research going forward as it allows to capture more gradual, fine-tuned maturation whilst taking individual differences into consideration.

Overall, the debate as to whether development in itself is a continuous (i.e. gradual) or discontinuous process (i.e. in stages) is not a new one, and has been widely discussed (e.g Lally & Valentine-French, 2019). However, this thesis is not set out to critique the merit of either approach. Instead, we argue that studying development with rigid chronological age groups is less beneficial when there is little consistency in definitions and when developmental stages include overlapping age ranges. Therefore, listener age will be investigated predominantly as a continuous variable throughout this thesis. Yet, age group analysis will be added to the developmental investigations to draw comparisons to previous research findings.

1.3 Aims of this thesis

Overall, much work remains to be done to understand how perceptions of vocal trustworthiness and emotion develop during the early lifespan. This thesis aims to expand on the existing literature and attempts to address the challenges that were identified in previous sections. Therefore, this thesis sets out to address three overarching goals: 1) address the gap in the literature in relation to the developmental trajectories of perceived vocal trustworthiness; 2) use socially-relevant word stimuli to investigate the developmental trajectories of perceived vocal emotion; and 3) create a vocal corpus pre-validated on social perceptions that is openly available to other researchers studying vocal personality and vocal emotion.

First, we address the current gap in the literature regarding developmental aspects of perceived vocal trustworthiness (Chapter 2). Recordings of the socially-relevant word “hello” will be used as it is a highly ecologically valid stimulus that has already been used with adult listeners (Baus et al., 2019; Mahrholz et al., 2018; McAleer et al., 2014). Analysis will focus predominantly on treating age as a continuous variable. Furthermore, to take individual differences into account, linear-mixed effects models including random-effects structures will be used to model relationships. Given there is no research on

potential gender differences in the developmental patterns of vocal trustworthiness perceptions, listener sex and speaker sex will be incorporated into the modelling approach.

Secondly, for the developmental trajectory of perceived vocal emotion, we identified that research has not included socially-relevant word stimuli. This is addressed in Chapter 3. Due to being interested in the fine-grained developmental aspects of emotion, age will foremost be analysed as a continuous variable. Similar to Chapter 2, mixed-effects modelling will be used, again including a random-effects structure. Listener sex and speaker sex will also be added to the model as variables of interest. Data will also be analysed treating age in age groups to be able to compare findings to previous research and identify similarities and differences between the two model approaches.

The third overarching goal for this thesis is the creation and development of an open-access database (Chapter 4) that includes emotive and neutral stimuli from the same speakers. Given the gap in the literature about the developmental aspects of vocal emotion recognition from socially-relevant words of high ecological validity, the corpus will start with the stimulus “hello”. This corpus is intended to serve as the baseline to further investigate the concepts of vocal personality and emotion perceptions either separately or simultaneously in future. Making the database available to other researchers (under a CC 4.0 license) will contribute to open science and reproducibility. Similar to Chapters 2 and 3, listener sex and speaker sex will be included in all statistical analyses.

The stimuli will be validated on a variety of personality and emotion ratings:

- The actor’s intended emotion and their intended intensity (as noted during the recording procedure);

- Ratings of perceived trustworthiness, dominance, and attractiveness (Study 1);

- Recognition accuracy and chance-corrected recognition rates (compared to intended emotion; Study 2);

- Recognisability, authenticity, and a composite score of both (Study 2);
- Valence and arousal ratings (Study 3);
- Perceived intensity (Study 4);
- Categorical labels from a free-labelling task[2];
- Acoustic measures[3].

[2] The data from the free-labelling task are currently being analysed, and is therefore not included in thesis chapter 4.

[3] It should be noted, that the collection of acoustic measures was described in Chapter 4 (but not analysed further).

Chapter 2 Development of perceived vocal trustworthiness across the early lifespan

2.1 Introduction

Voices convey socially relevant information about a speaker, such as their age (Hughes & Rhodes, 2010; Krauss et al., 2002; Moyse et al., 2014), gender (Schvartz & Chatterjee, 2012), race (Kushins, 2014), physical attributes (Pisanski et al., 2014; Pisanski et al., 2016; Sell et al., 2010), emotional states (Banse & Scherer, 1996; Pell & Skorup, 2008), and personality traits (McAleer et al., 2014; Oleszkiewicz et al., 2017). Whether they are accurate in relation to a person's true personality is questionable, yet, those quick judgements inform our behaviours towards a person. Recognising this information is essential for sustaining our wellbeing through identifying other's intentions, avoiding threats, and establishing meaningful relationships (Castle et al., 2012; Hawkley & Capitanio, 2015; McAleer et al., 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Zebrowitz & Montepare, 2008).

A proposed key aspect to developing and maintaining happy well-functioning relationships is trust (Simpson, 2007). Trustworthiness has been established by principal component analysis as one of the two main underlying dimensions (the other one being dominance) of person perception across research fields (Fiske et al., 2007; McAleer et al., 2014; Oosterhof & Todorov, 2008; Sutherland et al., 2013). Research has also suggested that adults form these first impressions of trustworthiness within brief exposure to the stimuli: 100 ms of exposure to a static face or listening to a short word such as "hello" (< 500 ms) is sufficient to form reliable judgements of how trustworthy a person appears (Mahrholz et al., 2018; McAleer et al., 2014; Oleszkiewicz et al., 2017; Willis & Todorov, 2006). In voice research, these first impressions have been shown to be highly consistent between listeners (Borkowska & Pawlowski, 2011; Klofstad et al., 2015; Klofstad et al., 2012; Latinus & Belin, 2011b; Mahrholz et al., 2018; McAleer et al., 2014; Rezlescu et al., 2015; Tigue et al., 2012), and stable within listeners across a variety of durations independent of speech content (Mahrholz et al., 2018). Hence, judgements of trust are proposed to rely on predictive coding of a present target in comparison to an internalised normative, akin to that previously established for voice identity (Latinus & Belin, 2012). However, when

this normative for person perception is established in the life span remains unknown.

Research on the developmental trajectories of vocal trustworthiness has focused on older adults thus far (e.g. Schirmer et al., 2019), whilst studies investigating early life span development are missing. In fact, in the vocal domain, there are very few studies to date addressing the maturation of personality traits between childhood and adulthood in general. For example, research on vocal attractiveness found that female adolescents (12-15 years) and adults (20-34 years) but not children (7-10 years) preferred the more attractive lower-pitched men's voices (Saxton et al., 2006). Investigating the critical developmental period between childhood and adolescence more closely, Saxton et al. (2009, 2013) showed the same preference for lower male voice pitch for older (ca. 14 years, varied slightly across the two studies) but not younger girls (ca. 12 years). Contrastingly, younger (ca. 12 years) but not older boys (ca. 14 years) preferred higher-pitched female voices which are frequently rated as more attractive by adult men (e.g. Borkowska & Pawlowski, 2011). Saxton et al. (2009, 2013) reasoned that the female voice lowers during adolescence which may be taken as an indication of sexual maturation by the older boys. Taken together, these results suggest perceived vocal attractiveness to change during adolescence when mate-selection becomes important. Yet, it is questionable whether vocal trustworthiness mirrors the development of attractiveness between childhood and young adulthood given that attractiveness maps onto the PCA-dimensions of trustworthiness and dominance differently depending on speaker sex (McAleer et al., 2014).

Since findings from auditory perception research tend to parallel results from face research (e.g. McAleer et al., 2014; Young et al., 2020; but Lavan, Burton, et al., 2019), it is worth reviewing the literature on the development of perceived facial trustworthiness. Trustworthiness based on emotionally neutral face stimuli develops early on in life. Six- to 8-month old infants show a preference towards trustworthy-looking faces (Sakuta et al., 2018), and children as young as 3 years old are able to make mean/nice judgements about a face (Cogsdill & Banaji, 2015). Recent studies employing economic game paradigms (Charlesworth et al., 2019; Ewing et al., 2015; Ewing et al., 2019) have proposed

that children start acting upon these trust perceptions around the age of 5, by making judgements about a person's behaviour and modifying their own behaviour towards them. When tested in low-cognitive tasks, such as a 2-alternative-forced choice paradigm, 3-4 year-olds reached adult-like consensus in their ratings of face trustworthiness, albeit with less consistency compared to 5-6 year-olds, 7-10 year-olds, and adults (Cogsdill et al., 2014). Employing an explicit judgement task with a 9-point Likert scale, Caulfield et al. (2016) had 5-, 7-, 10-year-olds, and adults rate trustworthy and untrustworthy face stimuli. Whilst all age groups could distinguish between trustworthy and untrustworthy faces, the 5- and 7-year-olds rated trustworthy faces as less trustworthy, and untrustworthy faces as more trustworthy compared to 10-year olds and adults respectively. There were no significant group differences between the 5- and 7 year-olds or the 10-year-olds and adults. Taken together, these results suggest that the sensitivity to detect facial trustworthiness emerges during early childhood, and reaches adult-like consensus around age 10. Yet, it is uncertain whether these results are mirrored in the developmental trajectories of the vocal domain.

Furthermore, exact developmental patterns are difficult to detect when participants are grouped into categories according to their chronological age. Statistically, dichotomising and grouping continuous variables may result in losing information or decreasing power (e.g. Altman & Royston, 2006). Conceptually, whilst dividing life into developmental stages to incorporate biological, cognitive, and/or socioemotional changes is meaningful, assigning concrete age ranges to those stages has been challenging (see 1.2.5 in the General introduction for a detailed discussion). To outline briefly, childhood is usually defined as a period between age 5 or 6 until 11 (Lally & Valentine-French, 2019), yet some sources consider anyone under 14 years a "child" (United Nations, 2021). Adolescence usually starts between 10 and 12 years of age and lasts until 18 to 21 (Santrock, 2020), followed by early adulthood from late teens/early twenties until late thirties (Santrock, 2020) or mid-forties (Lally & Valentine-French, 2019). These broad suggestions leave room for flexibility when defining age groups and have resulted in considerable variability in the literature. Despite its challenges and limitations, the majority of research discussed opted to investigate developmental patterns between age groups.

The current study therefore addresses the caveat within the literature by investigating the maturation of perceived vocal trustworthiness treating listener age as a continuous variable. Mixed-effects model analysis including random structures for by-subject and by-item terms allows us to observe detailed developmental patterns that may occur between childhood and adulthood. Given the design similarities to Caulfield et al. (2016), we expect "fine-tuning" towards adult-like consensus to be reflected in the slope. Since the majority of literature implied sufficient ability to perceive facial trustworthiness at 5 years of age, the listener age effect of the vocal domain might be of small magnitude. Additionally, to capture the perception patterns across all vocal stimuli, correlation analysis is performed on three listener age groups based on chronological age. Given the similarities between the visual and auditory modality and the dissimilarities between trustworthiness and attractiveness in the vocal domain, we expect vocal trustworthiness to emerge during early childhood. Thus, we hypothesise strong positive correlations between the perceived vocal trustworthiness of children (5-10 years), adolescents (11-19 years) and young adults (20-29 years).

2.2 Methods

2.2.1 Ethics

All procedures were approved by the University of Glasgow College of Science and Engineering Ethics Committee (No. 300150157) and are in accordance with the ethical standards of the 1964 Declaration of Helsinki. All participants/ caregivers of participants provided consent after acknowledging their participation would be voluntary, their data stored and treated anonymously, and that they could withdraw at any time.

2.2.2 Participants

Initially, 282 participants were recruited for the study at the Glasgow Science Centre and the University of Glasgow via convenience sampling. Given the low recruitment numbers for participants 30 years and above, this study focuses on children, adolescents, and young adults. Listeners between 5 and 10 years of age are grouped as children and 11- to 19-year-olds as adolescents (as suggested by

Santrock, 2020; WHO, 2022a). The age range of the young adult listeners was chosen to match the age range of young adult speakers (20-29 years). The final sample size for this study is therefore 183 participants (95 females, 88 males, mean age = 15.4 ± 6.93 years, age range = 5-29 years). See Table 1 for a split by listener sex and age group.

Table 1: Demographics, separated by Listener Sex and Listener Age Group

Age Group	Listener Sex	N	Mean Age	SD Age	Age Range
Children	Female	29	8.2	1.40	5-10
Children	Male	30	8.4	1.25	6-10
Adolescents	Female	36	14.5	2.75	11-19
Adolescents	Male	29	13.0	2.31	11-19
Young Adults	Female	30	23.8	2.33	20-29
Young Adults	Male	29	24.9	2.68	20-29

Note. N = Number of listeners. Children (5-10 years), Adolescents (11-19 years), Young Adults (20-29 years).

2.2.3 Stimuli

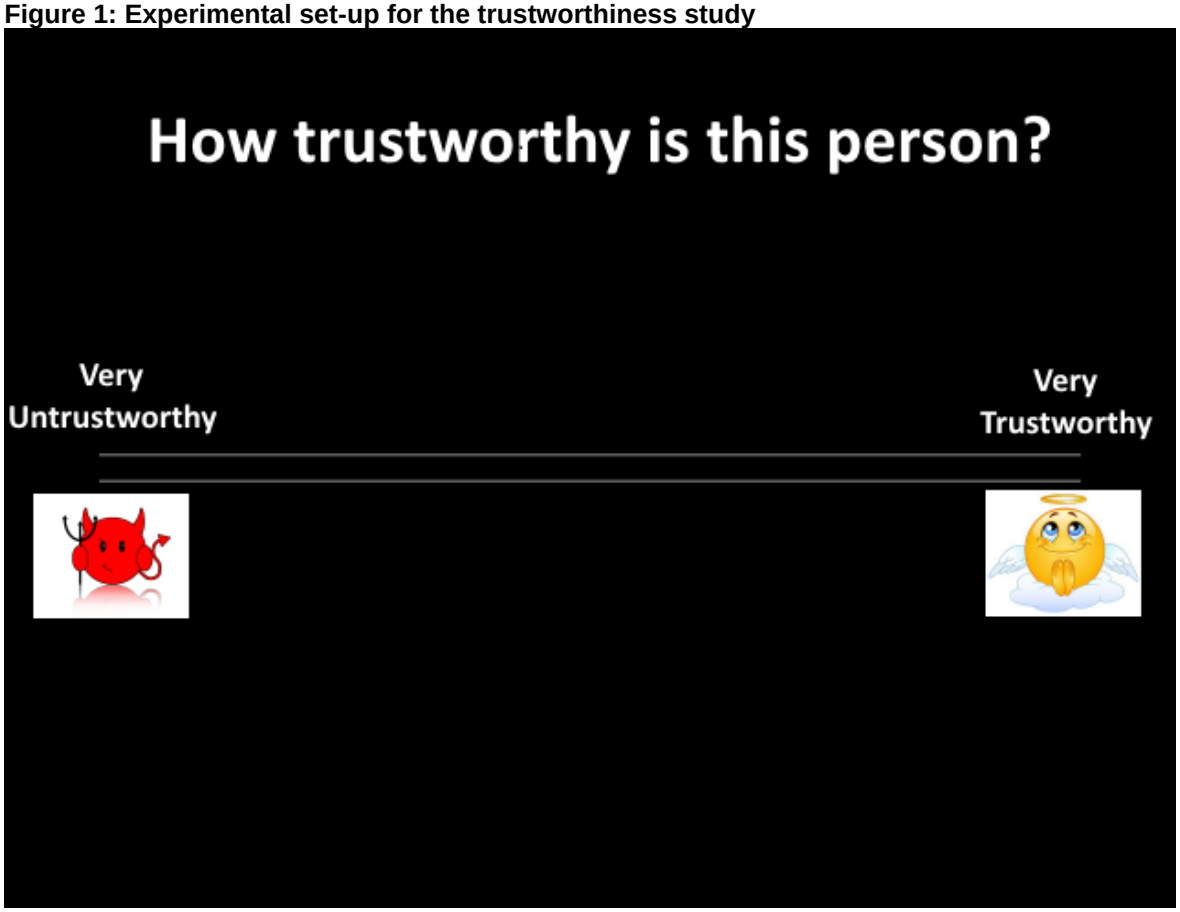
Voice recordings of 15 female (mean age = 22.9 ± 2.71 years) and 15 male native English speakers (mean age = 23.7 ± 2.79 years) saying the word “hello” were selected from the stimulus set used by McAleer et al. (2014). Stimuli were chosen from the full trustworthiness continuum of pre-ratings and controlled for height, weight, and age across speaker sex. Briefly, voices were recorded digitally (16 bit mono, 44100 Hz, WAV format) in an anechoic recording chamber and then normalised for power (RMS), and loudness using MATLAB (Version 9.1 (R2016b); MATLAB, 2016). Full voice recording and stimuli extraction procedures are reported in McAleer et al. (2014). The female and male voice stimuli had an average duration of 388.5 ms (SD = 70.52 ms) and 381.3 ms (SD = 50.09 ms) respectively. A Welch Two-Sample t-test showed no significant difference in stimulus duration between female and male speakers ($t(25.26) = 0.322$, $p = .751$).

2.2.4 Procedure

The experiment was conducted in the Glasgow Science Centre and the Psychology department of the University of Glasgow. Participants were asked to

provide written consent before starting the experiment. For participants under the age of 16, consent was provided by a parent or the primary caregiver. Participants received standardized example definitions and explanations of trustworthiness prior to testing (i.e. “by trustworthy we mean friendly, warm, honest”). They listened to both male and female stimuli in two consecutive blocks (1 block per speaker sex, in a counterbalanced order with a break in between) on a laptop via headphones and provided ratings on a visual analogue scale (VAS) slider with the endpoints “very untrustworthy” (left) and “very trustworthy” (right). The slider extremes were accompanied with age-appropriate emojis (see Figure 1) to facilitate understanding. After the voice was played, the VAS slider appeared until participants logged their decision by mouse click, followed by a 1000 ms break before the next trial. Each voice stimulus was presented twice in each block, resulting in a total of 60 ratings per participant. The study took approximately 10 minutes to complete.

Figure 1: Experimental set-up for the trustworthiness study



2.2.5 Sensitivity power analysis

We used package faux (Version 1.1.0; DeBruine, 2021) for data simulation to conduct a sensitivity power analysis for the main effects of listener age and presentation (i.e. first or second rating). Effect sizes for all fixed and random terms, as well number of participants (95F, 88M), stimuli (15F, 15M), presentations (2 ratings per stimulus), and distribution parameters for continuous age were extracted from the computed linear mixed-effects model. Repetitions were set to 10,000 and alpha to .05. For listener age, the minimum detectable effect size is 0.180 to achieve power of .8, and 0.209 for power of .9. For presentation, the smallest effect size detectable is 1.23 for power of .8, and 1.43 for power of .9.

Sensitivity power calculations for the correlation analysis were computed using the pwr package (Version 1.3.0; Champely, 2020) in R (Version 4.0.4; R Core Team, 2020). With 15 speakers per speaker sex, an assumed power of .8, and alpha of .05, the minimum detectable correlation coefficient is .657. For an assumed power of .9, the smallest correlation coefficient that can be detected is .722.

2.3 Results

2.3.1 Initial data preparation and analysis

All data wrangling, visualisations, and analysis was completed in R (Version 4.0.4; R Core Team, 2020) and RStudio (Version 1.1.463; RStudio Team, 2016). The following packages were used: car (Version 3.0.10; Fox & Weisberg, 2019), tidyverse (Version 1.3.1; Wickham et al., 2019), ggpubr (Version 0.4.0; Kassambara, 2020), irrNA (Brückl & Heuer, 2021), lme4 (Version 1.1.26; Bates et al., 2015), optimx (Version 2020.4.2; Nash, 2014; Nash & Varadhan, 2011), emmeans (Lenth, 2021), broom.mixed (Version 0.2.6; Bolker & Robinson, 2020), performance (Version 0.8.0; Lüdecke et al., 2021), faux (Version 1.1.0; DeBruine, 2021), pwr (Version 1.3-0; Champely, 2020), and Hmisc (Version 4.5.0; Harrell Jr, 2021).

Trustworthiness was operationalised from 0 to 100 on a VAS scale. All participants completed the experiment, hence no data were excluded. For

correlation analysis, trustworthiness ratings were z-scored for each participant to account for individual scale use, and an average z-scored trustworthiness value was calculated for each vocal stimulus in each listener age group. For any mixed-effects model analysis, raw scores were used since the random-effects structure accounts for individual scale use, and presentation (i.e. first or second rating) was added as a control variable. In the exploratory analysis of scale use, the difference between highest and lowest VAS value was computed for each participant.

2.3.2 Inter-rater consistency

To establish a level of consensus between listeners for each age group, intraclass correlation coefficient (ICC) estimates and their 95% confident intervals were computed on the z-scores trustworthiness ratings (averaged across the two presentations for each stimulus). Given participants rated both male and female vocal stimuli, analysis was performed across speaker sex. Computations were based on average ratings, type consistency of a two-way mixed-effect model (Hallgren, 2012; Koo & Li, 2016; Shrout & Fleiss, 1979). Interrater reliability can be considered excellent for all three age groups (Koo & Li, 2016; see Table 2), which is comparable to Cronbach's alpha values reported elsewhere (Mahrholz et al., 2018; McAleer et al., 2014; Oosterhof & Todorov, 2008; Willis & Todorov, 2006).

Table 2: Intraclass correlation coefficients with 95% confidence intervals for perceived trustworthiness of children, adolescents, and young adults

Listener Age Group	ICC	CI Lower Boundary	CI Upper Boundary
All Listeners (5-29 years)	.987	.979	.993
Children (5-10 years)	.927	.884	.960
Adolescents (11-19 years)	.973	.957	.985
Young Adults (20-29 years)	.971	.954	.984

2.3.3 Development of perceived vocal trustworthiness

We computed a linear mixed-effect model with by-subject random intercepts and slopes and by-item random intercepts to investigate the developmental course of perceived trustworthiness across listener age. By including listener sex

and listener age to the by-item, and presentation to the by-subject random effects structure, the model was overfitting. The terms were henceforth removed which resulted in an otherwise maximal model (see model (1) below). The dependent variable was the raw trustworthiness ratings. Listener sex, speaker sex, and presentation were added as control variables. Listener age was centred, and listener sex, speaker sex, and presentation were deviation-coded. The optimizer optimx (Nash & Varadhan, 2011) was used to avoid conversion issues.

(1) Trustworthiness ~ Listener Age + Listener Sex + Speaker Sex + Presentation +
Listener Age:Listener Sex + Listener Age:Speaker Sex +
Listener Age:Presentation +
(1 + Speaker Sex | Listener ID) + (1 | Speaker ID)

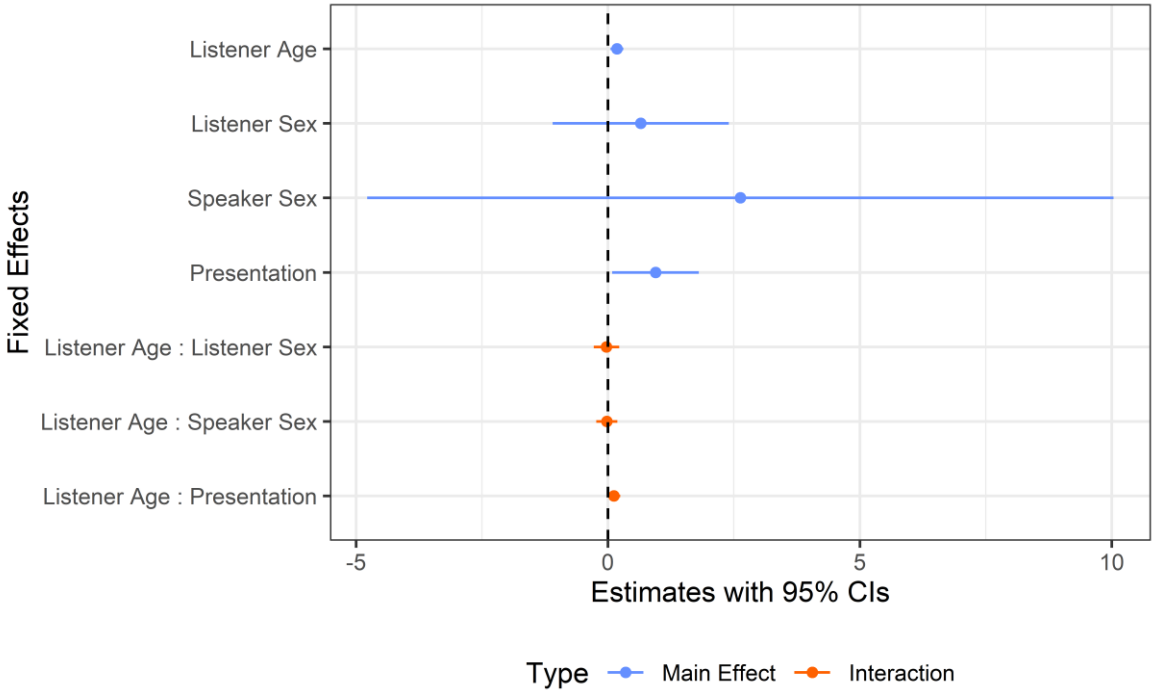
The model returned a main effect of listener age ($X^2(1) = 7.94$, $p = .005$). Perceived trustworthiness increased with increasing listener age by 0.18 VAS units ± 0.06 (standard error), meaning that with each additional year of age, listeners perceive the speakers as slightly more trustworthy. The model also returned a main effect of presentation ($X^2(1) = 4.64$, $p = .031$) showing that listeners across age rated a novel vocal stimulus as more trustworthy in their first presentation. On average first ratings were 0.95 VAS units higher than second ratings. There were no further significant main effects or interactions (all $p > .05$). See Table 3 for a model summary of all main effects and interactions and Figure 2 for the forest plot.

Table 3: Model summary of the linear mixed-effect model for all main effects and interactions

Term	Estimate	SE	t	LCI	UCI	Type
Listener Age	0.181	0.06	2.82	0.06	0.31	ME
Listener Sex	0.652	0.89	0.74	-1.08	2.39	ME
Speaker Sex	2.63	3.63	0.72	-4.49	9.75	ME
Presentation	0.945	0.44	2.15	0.08	1.81	ME
Listener Age:Listener Sex	-0.028	0.13	-0.22	-0.28	0.22	Int
Listener Age:Speaker Sex	-0.021	0.11	-0.20	-0.23	0.19	Int
Listener Age:Presentation	0.119	0.06	1.87	-0.01	0.24	Int

Note. **SE = Standard Error; CI = Confidence Interval; ME = Main Effect; Int = Interaction.**

Figure 2: Forest plot for main effects and interactions included in the linear mixed-effect model



Note. **Estimates are the fixed effects with 95% Confidence Intervals.**

The model returning an effect size of 0.181 is slightly above the smallest effect size detectable (i.e. 0.180 for power of .8) that was determined during sensitivity power analysis. This suggests that the study is sufficiently powered for the main effect of listener age. However, for presentation, the smallest effects that could be reliably detected is larger than the effect size obtained during analysis. This means there is uncertainty whether a true effect exists or if findings are due to chance. This should be investigated in future with a higher-powered design.

2.3.4 Patterns of perceived vocal trustworthiness

The analysis above shows the developmental trajectory of vocal trustworthiness across age, however, it does not provide an insight into perception patterns for individual vocal stimuli. Therefore, we conducted correlation analysis between the three age groups. Since listener sex was not a significant predictor of perceived trustworthiness, analysis was conducted regardless of listener sex (see also Amir & Levine-Yundof, 2013; Bruckert et al., 2010; Mahrholz et al., 2018).

All assumptions for Pearson correlation analysis were met. Correlation coefficients were very strong and positive between all three age groups for

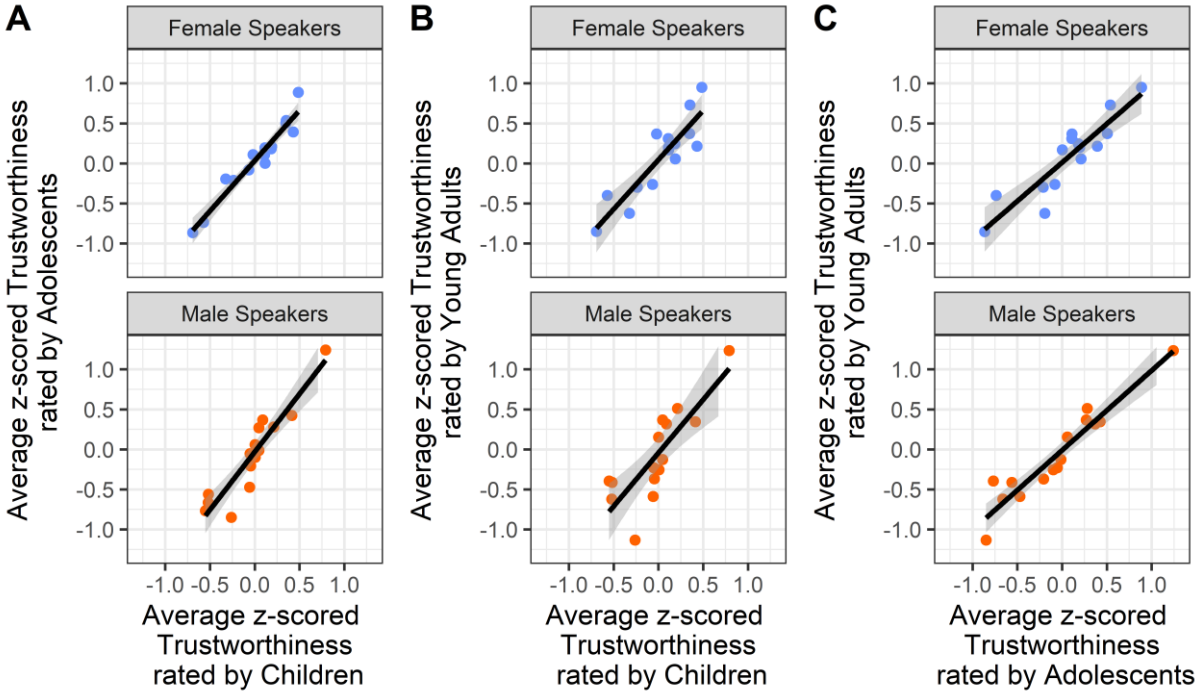
female as well as male voice stimuli (see Table 4). Figure 3 shows the scatterplot of z-scored mean trustworthiness ratings between all age groups in female and male voices respectively. It may also be worth highlighting that the slopes of the lines of best fit are visually steeper in the scatterplots when children's responses are correlated with adolescents and young adults compared to the adolescent/ young adult scatterplot. This suggests children perceive lower-trustworthy voices as more and higher-trustworthy voices as less trustworthy compared to adolescents and young adults.

Table 4: Pearson correlation coefficients and p-values for the listener age groups per speaker sex

Listener Age Group Comparisons	Female Speakers		Male Speakers	
	Pearson's r	p-value	Pearson's r	p-value
Children – Adolescents	.967	≤ .001	.929	≤ .001
Children – Young Adults	.892	≤ .001	.818	≤ .001
Adolescents – Young Adults	.907	≤ .001	.953	≤ .001

Note. **Children (5-10 years), Adolescents (11-19 years), Young Adults (20-29 years)**

Figure 3: Scatterplot of average z-scored trustworthiness ratings in female (top row) and male (bottom row) voices between Children and Adolescents (A), Children and Young Adults (B), and Adolescents and Young Adults (C)



Note. **Each point represents a Speaker ID. Children (5-10 years), Adolescents (11-19 years), Young Adults (20-29 years).**

The correlation values of .818 and above are larger than the minimum correlation values detectable computed in the sensitivity power analysis (.657

for power of .8, and .722 for power of .9). This means despite having a small number of vocal stimuli, the correlation analysis is sufficiently powered.

2.3.5 Exploratory analysis of the differences between first and second ratings

Despite having a main effect of presentation that may potentially be underpowered, we decided to explore the consistency of trustworthiness ratings across listener age by analysing differences between listeners' first and second ratings for each voice. We computed a linear mixed-effect model with by-subject random intercepts and slopes and by-item random intercepts. The dependent variable was the absolute value of the raw difference trustworthiness score between first and second ratings that were to be predicted from listener age. Listener sex, speaker sex, and interactions of each term with listener age were added as control variables. Listener age was centred, and listener sex, speaker sex, were deviation-coded. Including listener sex and/or listener age to the by-item random structure overfitted the model and were therefore removed. The final maximal model (2) is shown below. The optimizer optimx (Nash & Varadhan, 2011) was used to avoid conversion issues.

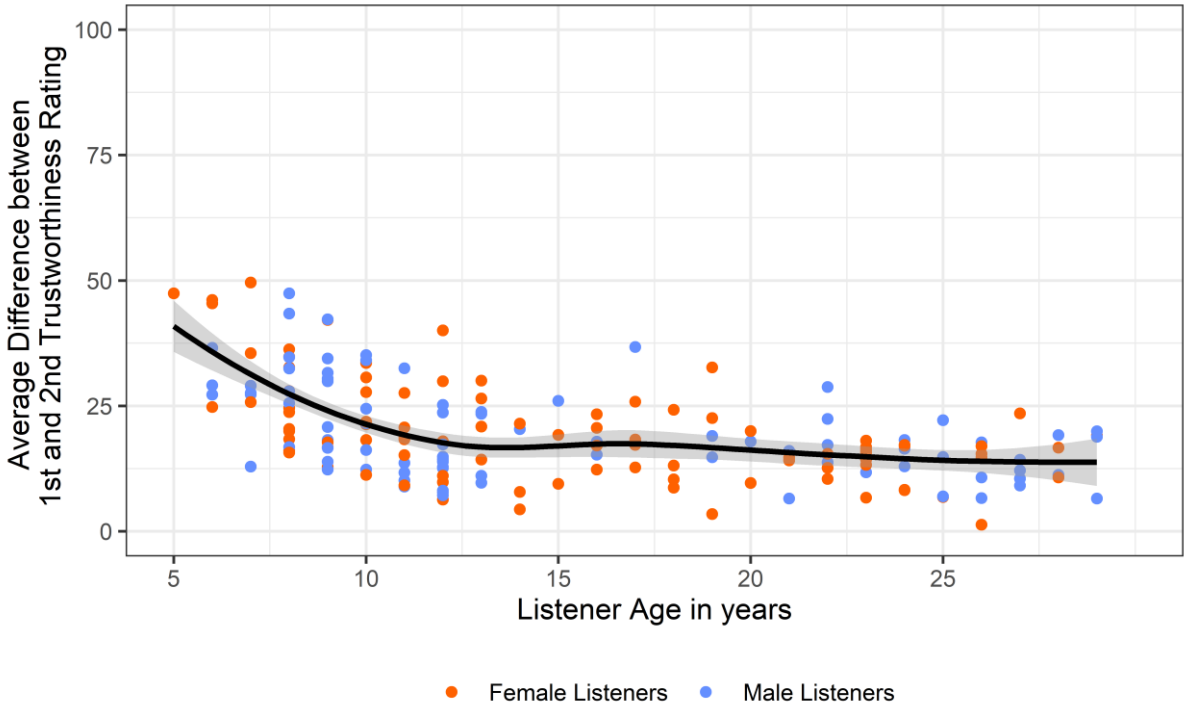
(2) Difference ~ Listener Age + Listener Sex + Speaker Sex +
Listener Age:Listener Sex + Listener Age:Speaker Sex +
(1 + Speaker Sex | Listener ID) + (1 | Speaker ID)

Assumptions were assessed visually and showed slight deviations from normally distributed residuals, and homoscedasticity. Schielzeth et al. (2020) addressed violations of distributional assumptions in linear mixed-effects models, and found estimates of interest to be generally robust. For violations of homoscedasticity, fixed effects were unbiased, yet confidence intervals were estimated too low. However, variances for normality violations were shown to be more variable and thus returning less precise estimates. This means that despite being relatively unbiased, our estimates might be further from the true value due to the violations of distributional assumptions.

Results showed a significant main effect of listener age on difference ratings (X^2 (1) = 60.59, $p < .001$). The difference between first and second rating was

predicted to be reduced by an average of 0.696 VAS units ± 0.089 (standard error) with each additional year of listener age. Whilst a linear relationship was found across the whole sample, plotting loess line of best fit seems to suggest a steeper decline of difference scores between childhood and adolescence (Figure 4). There were no further significant main effects or interactions. These results imply that trustworthiness perceptions become more consistent and "fine-tuned" with age, potentially with the main developmental course between childhood and adolescence.

Figure 4: Difference between listeners' first and second trustworthiness rating for continuous listener age



Note. **Each point represents a listener's difference score. Loess line of best fitted was added to capture finer detail of developmental patterns.**

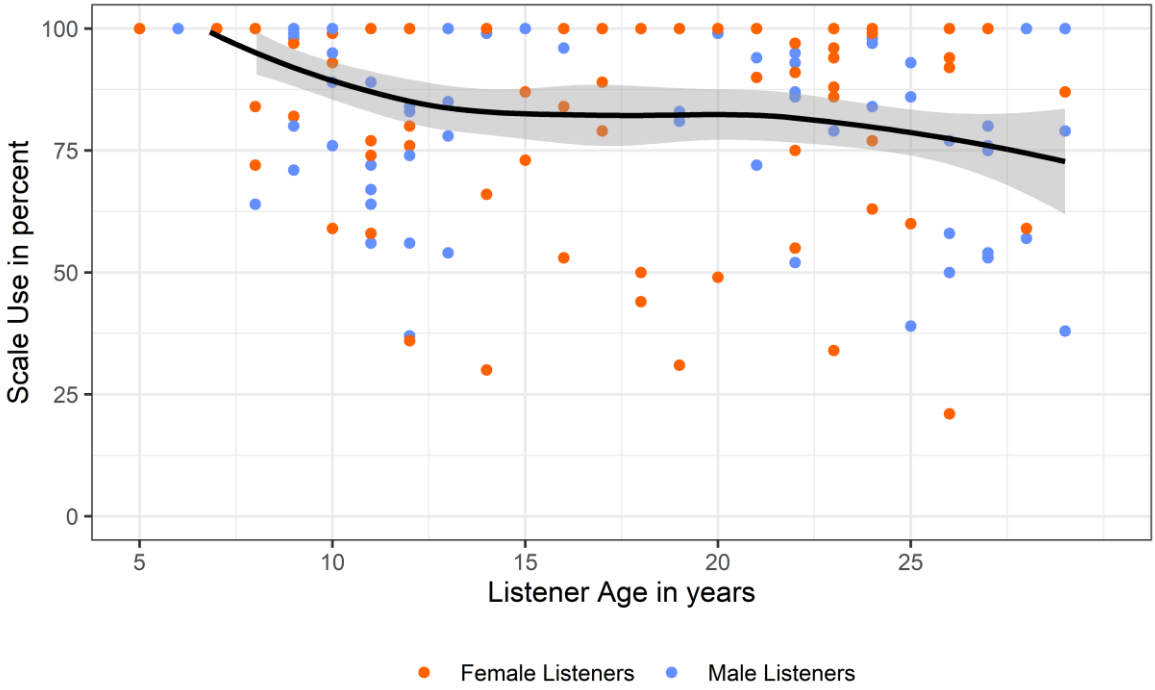
2.3.6 Exploratory analysis of scale use

The data were also explored to analyse whether listener age could predict how much of the rating scale was used. Therefore, the difference between the highest and lowest trustworthiness rating across all 60 ratings was determined for each participant. Since the VAS slider ratings in the experiment were recorded as numeric values from 0 to 100, the scale use is presented as a percentage. A linear model was computed predicting scale use from listener age, listener sex, and interaction terms between them. Listener age was centred, and deviation coding was used for listener sex. As above, slight deviations from

normality and homoscedasticity existed suggesting estimates may be further from true values than predicted by our model.

Results showed a significant main effect of listener age on scale use ($F(1) = 22.118$, $p < .001$; see Figure 5). With each additional year of age, scale use was predicted to decrease by 0.906 VAS units ± 0.193 (standard error). There was no significant main effect of listener sex or the interaction of listener age and listener sex. The results may suggest that at a young age, we see the world in "black-and-white" and develop a more "nuanced" scale use with increasing age.

Figure 5: Scatterplot of listeners' scale use for continuous listener age



Note. **Each point represents a listener's value of scale use. Loess line of best fitted was added to capture finer detail of developmental patterns.**

2.4 Discussion

The purpose of this study was to determine the developmental course of perceived vocal trustworthiness between childhood and early adulthood. We investigated whether listener age could predict trustworthiness perceptions by applying linear mixed-effects modelling, and found a small yet statistically significant increase of perceived trustworthiness with listener age. We hypothesised and found very strong correlations of vocal trustworthiness perceptions between all three age groups (children, adolescents, and young adults). In an exploratory approach, we showed that both the difference

between 1st and 2nd rating and the range of scale use decrease significantly with an increase in listener age. These results suggest that vocal trustworthiness perceptions are already existent by age 5 but become more "fine-tuned" and "nuanced" with increasing age.

The study's main aim was to investigate the developmental trajectory of perceived vocal trustworthiness. The majority of published papers investigating age differences have allocated participants to pre-determined age groups (Caulfield et al., 2016; Cogsdill et al., 2014; Ewing et al., 2015). However, there are drawbacks of dichotomising and grouping, for example losing information or decreasing power (e.g. Altman & Royston, 2006). We have addressed this caveat of the literature by treating age as a continuous variable and employing linear mixed-effects model analysis including fixed as well as random structures. We found a small main effect of listener age implying that participants' perception of vocal trustworthiness increased slightly with an increase in age. However, the trustworthiness ratings provided by children, adolescents, and young adults were highly correlated between all three age groups in both female and male voices. Additionally, we obtained inter-rater consistency between listeners within each age group as well as across all listeners which is consistent with reports in adult populations (Lortie et al., 2018; McAleer et al., 2014; Oosterhof & Todorov, 2008; Rezlescu et al., 2015). Taken together, these results suggest that a gauge of trustworthiness is already present within children around 5 years of age. Regardless of their age, listeners agreed which voices sound trustworthy and which ones less so. This may suggest that at some point before the age of 5, listeners develop an internalised prototypical representation of trustworthiness usually observed in adults (Latinus & Belin, 2012).

The developmental trajectory of perceived vocal trustworthiness mirrors patterns observed with face research employing similar judgement and decision-making paradigms (Caulfield et al., 2016; Charlesworth et al., 2019; Cogsdill et al., 2014), suggesting the perception of trustworthiness is modality-independent. For example, Cogsdill et al. (2014) reported a main effect of age group when rating facial trustworthiness in a 2-AFC paradigm, however post-hoc analysis revealed differences to be significant only between the age group of 3-4 year-olds and all other age groups; by age 5-6, adult-like consensus was reached. In

contrast, Caulfield et al. (2016) showed that 5- and 7-year olds already possessed the ability to evaluate trustworthiness when judging face stimuli on a more complex Likert rating paradigm, yet adult-like consensus was only reached by 10 years of age. These findings propose that adult-like consensus may be reached by age 5 when utilising easier rating paradigms that require less cognitive load (such as the 2-AFC by Cogsdill et al., 2014). However, when employing more complex paradigms, subtle developmental patterns may become apparent. Using a similar rating paradigm to Caulfield et al. (2016), our small yet gradual increase in trustworthiness could be reflected in their group differences found between 7 and 10 years of age. This again emphasizes the necessity to use continuous age in future research to be able to capture detail of gradual development rather than assuming a sudden change at a specific age.

One interesting finding observed visually concerns the lines of best fit in the scatterplots (Figure 3). When displaying the relationship of perceived vocal trustworthiness between children and adolescents or children and young adults, the slopes appear steeper compared to the scatterplot showing the relationship between adolescents and young adults. This suggests that whilst high- and low-trustworthy voices are perceived as such by all listeners, children's perception seemed to be centred towards the mean. This can also be seen in Caulfield et al. (2016) showing the difference scores between trustworthy- and untrustworthy-looking faces to be significantly smaller in younger children compared to older children and adults. Here, we observe a similar pattern despite selecting stimuli from a continuum of vocal trustworthiness rather than from the extreme ends of the pre-validated dataset (McAleer et al., 2014). Potential explanations might lie within the consistency between first and second rating or the scale use since exploratory analyses showed that both the difference between the two ratings and the percentage of scale use decreased with an increase of listener age. However, the scatterplots of Figure 3 should have accounted for the differences of scale use by depicting z-scores. This implies that consistency of those perceptions increases or becomes gradually more "fine-tuned" and more "nuanced" over time. Yet, since this analysis was exploratory, future research should investigate with a more suitable paradigm.

One limitation in this study might be that vocal stimuli were provided by young adult speakers between 20 and 29 years of age. It could be argued that this may have lead to an "own-age bias" suggested in some age groups when studying vocal emotion (Amorim et al., 2021) or in research including face stimuli (Hills & Willis, 2016; but see Griffiths et al., 2015). However, a bias cannot be observed within the present study since ICC values were consistently high within all listener groups as well as across all listeners. The very strong correlation values between all three age groups further support the notion that children, adolescents, and young adults perceive trustworthiness equally in the selected voice sets. Our results of no "own-age bias" are in concordance with research on trustworthiness using facial stimuli from children and adults (Ewing et al., 2019) or younger and older adult's vocal stimuli (Schirmer et al., 2019). However, since we only presented vocal recordings from young adult speakers, we cannot make concrete claims within this study. Future studies should specifically investigate with vocal recordings from children, adolescents, and adults.

A second limitation may be the inclusion of the angel and devil emojis on the rating scale of trustworthiness. Whilst this was done to remind the youngest participants of the rating scale anchors during the task (see Mackenzie et al., 2018 for the benefits of using emojis with children), it may have altered the concept of trustworthiness in children as compared to adults who could be relying more on the labels (Herring & Dainas, 2020). For example, participants may have interpreted the angel as "good" or "nice" and the devil as "bad" or "mean". When making mean/nice judgements, 5- to 6-year-olds have been shown to rate faces similarly to adults (Cogsdill & Banaji, 2015; Cogsdill et al., 2014). If this were also the case in the vocal domain, it may be an alternative explanation for children's adult-like perception ability in this study. To investigate further, this study should be replicated without any emojis or with different kinds of emojis on the rating scale.

One avenue for future research would be the investigation of the slight positivity effect observed within this study. Usually, an age-related positivity effect is reported to emerge in middle and late adulthood when people start favouring positive over negative stimuli (Reed & Carstensen, 2012). In face research, older adults have been shown to rate either all stimuli (Zebrowitz et al., 2017) or the

least trustworthy-looking ones (Bailey et al., 2015) as more positive compared to younger adults. Schirmer et al. (2019) did include listener age in their mixed-effects model analysis and reported “a marginal Listener Age effect” (p. 7) showing trustworthiness perceptions by older adult listeners to be more positive than those by younger adults. Since the effect was only marginal, it may imply small effects that should be studied more explicitly in future. It would indeed be of interest to investigate whether this gradual positivity trend found in this study continues into middle and late adulthood or whether it plateaus past age 29 before re-emerging in middle to late adulthood. Analysis would benefit treating listener age as a continuous variable to observe detailed patterns in the developmental trajectories.

In conclusion, our results contribute to the field of personality research by showing how vocal trustworthiness develops and matures between childhood and young adulthood. We have demonstrated that reliable perceptions of trustworthiness are already present during childhood, but ratings become slightly more positive as people age. This could indicate a potential positivity effect to emerge much earlier than during the proposed middle and late adulthood. However, we have also shown that perceptions become more “fine-tuned” and “nuanced” with increasing age.

Chapter 3 Development of perceived vocal emotion across the early lifespan

3.1 Introduction

The human voice is a rich source of social information, not only conveying linguistic cues, such as a speaker's intentions and beliefs through explicit verbal content, but more importantly non-verbal para- and extralinguistic cues (Laver, 1994; Schweinberger et al., 2014). Extralinguistic features reveal more stable physical characteristics about a speaker such as age (Demenescu et al., 2014; Lima et al., 2014; Lortie et al., 2018), sex (Schvartz & Chatterjee, 2012), race (Kushins, 2014), identity (Lavan, Knight, et al., 2019), and personality traits (Aronovitch, 1976; Borkowska & Pawlowski, 2011; McAleer et al., 2014; Oleszkiewicz et al., 2017), whereas paralinguistic features, such as the tone of voice, provide insight to a speaker's affective state (Banse & Scherer, 1996; Lima et al., 2013). Research has shown that we extract these cues to form social impressions after very brief exposure to the voice (McAleer et al., 2014), that judgements are consistent across listeners and reliable across short durations of varying content (Mahrholz et al., 2018), and that they can be made across languages irrespective of the listener's own language ability (Baus et al., 2019). Whether accurate or not, they subsequently influence decisions we make towards the speaker, such as mate choices (Apicella & Feinberg, 2009) or voting preferences (Klofstad et al., 2015). However, it is as important to focus on paralinguistic information (i.e. how something is said) as it is to pay attention to content (i.e. what is being said). The majority of interpersonal communication is conveyed through non-verbal means (Burgoon et al., 2010) and paralinguistic cues could place emphasis or alter the meaning of the message being communicated (e.g. Rivière & Champagne-Lavau, 2020; Voyer et al., 2016). In particular, assessing emotional states in others quickly is an essential part of social communication that facilitates more meaningful interactions, helps to communicate effectively, and, on a more basic level, allows to identify whether to approach or avoid a person (Kamiloğlu et al., 2020; Litt et al., 2020; Mennella et al., 2020; Scherer, 2009).

Darwin (1872) was one of the first to discuss how certain emotion categories play an essential role in survival. Whilst anger and fear both alert to threats, anger

motivates confrontational approach behaviour, whereas fear promotes withdrawal (Darwin, 1872; Fessler et al., 2004; Lerner & Keltner, 2001; Moons et al., 2010; Shariff & Tracy, 2011). Similarly, experiencing disgust warns of aversive foods or distasteful ideas and behaviours and triggers avoiding behaviours, such as pathogen avoidance (Cepon-Robins et al., 2021; Darwin, 1872; Shariff & Tracy, 2011). Building up from this approach, emphasizing the innate psychological and biological components of facial expressions, Ekman and Friesen (1971) created the "Basic Emotion Theory". Their influential theoretical framework proposes there are six basic overarching "emotion families" – namely happiness, sadness, anger, surprise, fear, and disgust – that are discrete and distinguishable from one another and universally recognised. Initially addressing face perception, these six basic emotion categories have been shown to replicate in voices research (Sauter, Eisner, Calder, & Scott, 2010) and across different cultures/ languages (Kawahara et al., 2017, 2021; Laukka & Elfenbein, 2021; Pell & Skorup, 2008; Sauter, Eisner, Ekman, & Scott, 2010).

Recently, research focused on investigating the early developmental course of emotion recognition from auditory stimuli has shown maturation between childhood and adulthood. Children as young as 5 years of age are able to recognise emotion at better than chance levels (Sauter et al., 2013; but see Aguert et al., 2013, for children age 5 performing at chance-level) with a steady increase in recognition abilities between the ages of 5 and 10 (e.g.Allgood & Heaton, 2015; Doherty et al., 1999; Sauter et al., 2013). When exactly adult-like performance is reached, remains an open question though: Some research (Brosgole & Weisman, 1995; Zupan, 2015) suggests adult-like accuracy being reached by early-adolescence (around 12 years of age). In contrast, Chronaki et al. (2015) and Tonks et al. (2007) found no significant differences between children and adolescents. Tonks et al. (2007) did not include an adult comparison group to draw conclusions as to whether adult-like consensus may be reached by 9 years of age. However, Chronaki et al. (2015) showed perceptions by both children and adolescents differed significantly from adults' suggesting that adult-like perception is reached after early adolescence. Grosbras et al. (2018) provided further support for maturation by middle adolescence, reporting that emotion recognition follows a quadratic pattern with steeper improvement between childhood and adolescence and adult-like performance levels reached

between ages 14 and 15. However, Morningstar, Ly, et al. (2018) found that adults had significantly higher accuracy for adult speech segments than 13- to 15-year-olds, suggesting adult-like consensus is more likely to be reached by late adolescence. Finally, Amorim et al. (2021) showed ongoing maturation between 15 and 23 year-old participants indicating development may continue beyond late adolescence or into early adulthood.

One reason for the sightly contradictory results could be that different emotion categories mature at different rates. For example, results from Morningstar, Ly, et al. (2018) were driven by significant differences in fear and sadness as mid-adolescents performance was on par with adults' perception for anger, happiness, and disgust. Slightly contrasting to Morningstar et al., Kawahara et al. (2017; observed from voice-only data provided by the authors) found that Japanese listeners' (age groups: 5-6, 7-9, 10-12, and adults) accuracy improved steadily with an increase in age for anger (74.78% to 92.42%), whilst happy Japanese speech segments were perceived consistently high (~90%) across age groups. These findings are partially in agreement with Grosbras et al. (2018) who showed happiness to be the most and anger the least well recognised category across age, however recognition accuracy of all emotion categories improved with increasing age (albeit anger the most). Investigating 10 emotion categories, Amorim et al. (2021) showed that pleasure, relief, sadness, and surprise were stable between childhood and adolescence, whereas all other emotion categories (achievement, anger, disgust, fear, happiness, and neutral) improved significantly. The significant changes between 15 and 23 years of age seem to be driven by fear and achievement with very little maturation of the other emotion categories. It can therefore be assumed that different emotion categories develop and mature at different rates. Since results are quite ambiguous as to which emotion categories drive the overall developmental changes observed in the existing literature, the current study aims to bring clarity by conducting a large-scale study of the basic 6 emotion categories.

Pinpointing when exactly adult-like emotion recognition manifests may also depend on specific stimuli types. So far, research has utilised either non-verbalised emotional interjections or affect bursts such as cries, laughs, and grunts (e.g. Allgood & Heaton, 2015; Amorim et al., 2021; Belin et al., 2008;

Chronaki et al., 2015; Laukka et al., 2013; Sauter & Scott, 2007), nonsense syllables (e.g. bábaba; Mildner & Koska, 2014), pseudo-words (Demenescu et al., 2014), three-digit numbers (Sauter et al., 2013), pseudo-sentences (Aguert et al., 2013; Banse & Scherer, 1996; Chronaki et al., 2018; Paulmann & Pell, 2010), sentences with neutral content (Kawahara et al., 2021; Morningstar, Ly, et al., 2018; Nelson & Russell, 2011; Zupan, 2015), or sentences employing congruous/ incongruous paradigms (e.g. happy content with happy/angry prosody; Friend, 2000; Morton & Trehub, 2001). Emotion recognition may mature earlier when less ambiguous stimuli are involved. For example, emotion recognition from non-verbal vocalisations is often seen as "easier" and results in higher recognition rates compared to emotion recognition from prosodic speech or incongruous paradigms which may require more cognitive load (Hawk et al., 2009; Hunter et al., 2010; Laukka et al., 2013; Nelson & Russell, 2011; Sauter et al., 2013). Developmentally, this can also be seen in lower improvement rates between younger (mean age 6 years 3 months) and older children (9 years 2 months) for non-verbal affect bursts compared to inflected speech of three-digit numbers (Sauter et al., 2013). Particularly with conflicting information between lexical content/context and prosody, children around age 5 years are thought to rely on content (Morton & Trehub, 2001) or contextual information (Aguert et al., 2013), and have difficulties switching between content and paralinguistic information (Morton et al., 2003; Waxer & Morton, 2011). There appears to be a shift at around 10 years of age when children start incorporating prosody for a more holistic emotion recognition approach (Friend, 2000).

Whilst emotion recognition from longer speech scenarios – especially incongruent paradigms – are applicable to daily social encounters, other stimulus types, such as three-digit numbers or syllables, are perhaps less so. Non-verbal affect bursts are very common in everyday speech, yet can still be of considerable duration (~1000 ms; e.g. Amorim et al., 2021; Belin et al., 2008; Sauter et al., 2013) which in turn could decrease ecological validity. There is a gap in the literature investigating emotion recognition from brief socially-relevant stimuli neutral in content. To our knowledge, only one study has utilised socially-relevant stimuli thus far (names of the children participating in the study; Mildner & Koska, 2014). However, the study only had three typically-developed children participating and no adult comparison group which in turn limits conclusions that

could be drawn from the sample. The field would benefit from using brief socially-relevant stimuli high in ecological validity, such as the word “hello”.

The literature has also reported an overall female advantage in recognising facial emotion, however gender differences for vocal emotion recognition results are still ambiguous. Whilst some studies report females to be more accurate than males in decoding emotion with varying stimulus types (Belin et al., 2008; Collignon et al., 2010; Grosbras et al., 2018; Keshtiari & Kuhlmann, 2016; Paulmann & Uskul, 2014; Sen et al., 2018), others found either differences of small magnitude (Lausen & Schacht, 2018; Thompson & Voyer, 2014), or see females’ and males’ accuracy in vocal emotion recognition on par (e.g. Amorim et al., 2021; Lima et al., 2014; Paulmann et al., 2008; Sauter et al., 2013). Similarly, when encoding information, speaker sex may influence accuracy ratings. Affect portrayed by female speakers seems better recognised than male speakers, for example when using affect bursts (Belin et al., 2008; Lausen & Schacht, 2018) and pseudo-words (Lausen & Schacht, 2018). However, Lausen and Schacht (2018) also found that males were better encoders for negative nouns, whilst no significant differences were reported for semantically positive or neutral nouns. The developmental trajectories of gender differences for listeners and speakers have thus far not been explored.

One final argument could be made critiquing grouping listeners by chronological age. Yet, the majority of papers in the field have done so which may have contributed to the substantial variability of developmental findings within the literature. There is remarkably little consistency in the literature regarding definitions and when developmental stages include overlapping age ranges (see Chapter 1 for a detailed discussion). Furthermore, using broadly defined age categories may result in the loss of detailed information or a decrease in power (e.g. Altman & Royston, 2006). Sauerbrei and Royston (2010) have argued that datapoints on either side of the grouping borders would be more similar to each other than they are different. Treating listener age as a continuous variable would be more beneficial for research going forward as it allows to capture more gradual, fine-tuned maturation whilst taking individual differences independent of chronological age group into consideration.

To summarise, this study aims to investigate the maturation of vocal emotion recognition between childhood and adulthood. Expanding on the stimulus types used in previous research, this study utilises recordings of the socially-relevant word ("hello") of around 500 ms which is more applicable to real-life situations with higher ecological validity. Since the literature has pointed to subtle developmental changes between certain age groups that may potentially get lost when grouping participants by age, this study investigates emotion recognition as a function of age as a continuous variable. However, data will also be analysed by dividing participants into age groups to allow observing similarities and differences in confusion patterns between childhood, adolescence, and early adulthood. Including all 6 basic emotion categories (happiness, sadness anger, surprise, fear, disgust), and a neutral representation permits to establish whether some emotion categories mature faster than others and which emotion categories drive overall recognition accuracy. Given the uncertainty as to whether a socially-relevant "hello" is semantically (dis)similar to neutral nouns or affect bursts, the current study will explore the developmental trajectories of gender differences for listeners and speakers in relation to recognition accuracy. Using an all-encompassing binomial mixed effect model with by-item and by-participant random intercepts and slopes, we hypothesise:

H1a: For age as a continuous variable, we expect emotion recognition ability for each emotion category to improve with increasing age.

H1b: Children will be significantly lower in emotion recognition ability compared to adolescents and young adults. Differences between adolescents and young adults will be smaller.

H2: There will be a significant two-way interaction between listener age and emotion category.

H3: Recognition accuracy will differ between female and male listeners. However, given the contradictory literature, we will refrain from a specific directionality for this hypothesis.

H4: Recognition accuracy will differ between female and male speakers. However, given the contradictory literature, we will refrain from a specific directionality for this hypothesis.

3.2 Methods

3.2.1 Ethics

Recording and experimental procedures were approved by the University of Glasgow Ethics Committee and are in accordance with the ethical standards of the 1964 Declaration of Helsinki (No. 300170216).

3.2.2 Power analysis

Power analysis was conducted during pre-registration using the power analysis tool PANGEA (https://jakewestfall.shinyapps.io/pangea/) designed by Jacob Westfall (Westfall, 2016). Assuming a power of .9 and alpha of .05, a minimum of 40 female and 40 male listeners within each age group for each speaker sex are required to achieve effect sizes of 0.31 (based on research by Lausen & Schacht, 2018; Ruffman et al., 2008).

3.2.3 Participants

Participants were recruited face-to-face within the Glasgow Science Centre (GSC). During the pandemic, advertising was placed on the University of Glasgow School of Psychology Subject Pool where participants were invited to take part in the experiment on the online platform pavlovia.org. In total 1039 participants were recruited (922 in GSC, 117 on Pavlovia). No monetary incentives were provided for partaking, however Level 1 Psychology students could receive 1 experimental credit as part of their course requirements.

Pre-stipulated criteria required us to remove 1) participants under the age of 5 to comply with ethical approval obtained (20 participants in GSC); 2) anyone who did not complete all 35 trials (abandoned or programme crashed) as an indication of withdrawn consent (27 in GSC, 2 on Pavlovia); 3) anyone who answered with one specific category 50% or more of all trials (13 in GSC); and 4) anyone who responded randomly in a clockwise or anti-clockwise pattern (1 in

GSC). Given the between-subject study design, in cases in which we could reliably identify that participants completed both male and female speaker experiments, only data from the first completed experiment was kept and the latter one was discarded (14 on Pavlovia). Furthermore, we removed those pre-defined age groups from this analysis for which the minimum recruitment number of 40 was not met (middle-aged adults aged 40 to 60 and older adults; 180 participants). The final sample size for the current study is therefore 782 (399 identified as female, and 383 as male; mean age = 17.1 ± 10.0 years, range: 5-39 years). For the age group analysis, participants were divided into children (5-10 years), adolescents (11-19 years), and young adults (20-39 years). The age groups of children and adolescents were chosen to align roughly with the existing literature (Santrock, 2020; Shaffer & Kipp, 2014; WHO, 2022a) and groupings defined in Chapter 2, whereas the adult age range was selected to approximately match the age range of the speakers in the Glasgow vocal emotion and personality corpus (Chapter 4). See Table 5 for a detailed summary per age group, and Supplementary material 1 for distribution of continuous age.

Table 5: Demographic profile of participants

Age Group	Listener Sex	N	Mean Age	SD Age	Age Range
Female Speaker Experiment					
Children	Female	75	7.4	1.6	5-10
Children	Male	65	7.6	1.6	5-10
Adolescents	Female	60	16.0	3.0	11-19
Adolescents	Male	60	16.0	3.1	11-19
Young Adults	Female	67	30.0	6.0	20-39
Young Adults	Male	63	28.6	6.7	20-39
Male Speaker Experiment					
Children	Female	74	7.8	1.6	5-10
Children	Male	70	7.9	1.6	5-10
Adolescents	Female	60	15.6	3.1	11-19
Adolescents	Male	66	15.4	3.4	11-19
Young Adults	Female	63	29.6	6.6	20-39
Young Adults	Male	59	28.6	6.8	20-39

3.2.4 Recording procedures and stimuli selection

Stimuli were selected from a pool of recordings (Chapter 4) provided by 24 native English speakers from Scotland (12 females) who were recruited for stimuli recording via the University of Glasgow School of Psychology Subject Pool. Advertising was placed for participants who have grown up in Scotland, under the age of 40 without speech impediments, and either have had experience in acting/ voice acting etc. or would be comfortable and confident to produce recognisable voice stimuli in a recording booth. These stimuli were part of a larger recording session, and speakers received £15 for their contribution.

During the recording sessions, speakers were told to emote the word “hello" in 6 affect categories (happy, sad, angry, surprised, fearful, disgusted), both in a low- and high-intensity representation, and in a neutral manner. To avoid confusion of the term intensity with loudness rather than an enhanced emotional attempt, the wording of “subtle” and “theatrical” was used during recording sessions to refer to low and high intensity respectively. No other instructions or vignettes were provided to allow speakers the freedom to express the emotions how they saw fit. The recordings took place in a custom-made sound-attenuated chamber within the School of Psychology at the University of Glasgow. Audacity Version 2.3.0 (Audacity Team, 2021) was used for recording, extracting, and editing the stimuli (.wav format, 16-bit mono, 44100 Hz). Within this paper, the neutral category will be treated as one of the affect categories (see also Amorim et al., 2021; Chronaki et al., 2015; Sen et al., 2018).

The stimuli were subsequently normalised for sound intensity (i.e. loudness) within each emotion category via MATLAB (Version 9.1 (R2016b); MATLAB, 2016). Normalising within each emotion category was done to adjust for sound intensity differences between speakers (for example: some speakers standing further away from the microphone than others), whilst maintaining information in relation to specific affect categories (Chen et al., 2012; Kamiloğlu et al., 2020; Schirmer et al., 2007).

The 312 stimuli (24 speakers x 13 vocal attempts) were pre-validated by three lab members on authenticity and recognisability measured on 4-point Likert scales ranging from 1 = “not at all recognisable/ authentic” to 4 = “very much

so". For each emotion category, the 5 stimuli with the highest average of the two scores were chosen as representations of their respective affect categories irrespective of speaker ID and intensity attempt. However, a speaker was only selected once per emotion category (in case both the low- and high-intensity representation were scored highly). This resulted in a total of 35 stimuli per speaker sex uttered by 10 female (mean age = 22.0 ± 1.8 years) and 12 male speakers (mean age = 22.3 ± 4.7 years). A Welch two sample t-test revealed no significant age differences between speakers, t(14.589) = 0.17, p = .866, d = 0.07. For practice trials, stimuli were chosen that were not represented in the experimental trials. Information on duration can be found below in Table 6.

Table 6: Acoustic information of mean duration and SD per Emotion, separately for stimuli encoded by female and male speakers

	Female Speakers		Male Speakers	
Emotion Category	Mean Duration	SD Duration	Mean Duration	SD Duration
Happiness	640.7	143.6	484.8	122.3
Sadness	521.7	72.9	520.9	63.8
Anger	558.1	85.8	571.3	135.2
Surprise	626.8	117.0	508.0	92.5
Fear	496.9	136.9	626.9	120.1
Disgust	635.4	119.8	718.4	223.3
Neutral	462.2	82.5	407.9	18.6

***Note*. All values in ms**

3.2.5 Procedure and experimental set-up

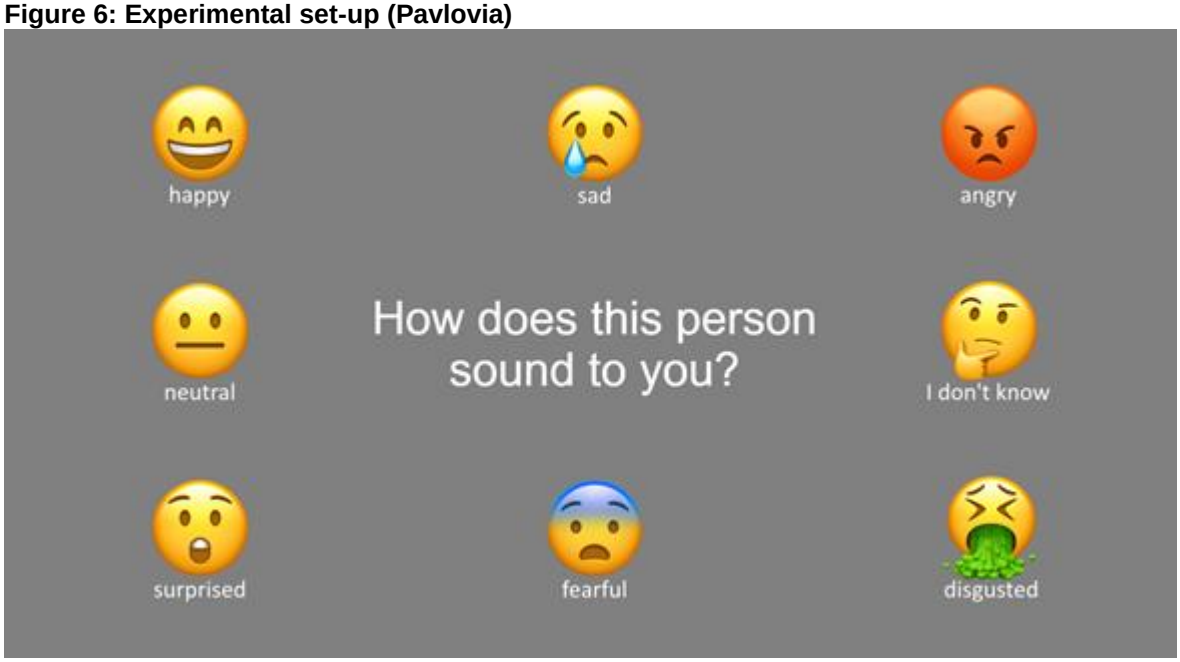
First, all participants were informed about the purpose of the experiment, and told their data would be contributing to a PhD in Psychology and a potential publication in an academic journal. They were also informed that their contribution would be voluntary, their data anonymised and securely stored, and that they would not have to complete the experiment if they did not wish to. For children, this was mostly delivered verbally with child-directed speech, whereas adult participants in the GSC were provided with an information form. For the online participants, information was presented within the experiment's webpage. Participants in the GSC completing the experiment in person provided written consent. Anyone under 16 years of age had a parent/ guardian/ caretaker providing consent. Participants taking part online via pavlovia.org (age

16+ years) were provided with a consent page on which they were informed that continuing to access the experiment would provide consent or in case they did not want to provide consent, instructions told them to press esc and close the browser. Participants were also made aware that closing the browser at any stage during the experiment would terminate participation and none of their already provided data would be used in the analysis.

Face-to-face participants completed the experiment on Samsung tablets (experiment created with OpenSesame Version 3.2.4; Mathôt et al., 2012), whereas online participants would be required to complete the experiment on their own computers via pavlovia.org (experiment created with PsychoPy v2020.2.4; Peirce et al., 2019). All participants provided demographic information about sex, age, nationality, and how long they have been in Scotland for. Additionally, for the online participants, an option was provided to indicate whether they wanted to receive participation credits for their contributions as part of their Psychology degree course requirements.

Participants were then shown an image of the experimental layout and told to choose the option that they think represents the emotion they hear best. There were 8 answering options in total: 6 emotion categories (happy, sad, angry, surprised, fearful, disgusted), a neutral option (when there is no emotion present), and an “I don’t know” option in case the emotion heard could not be sorted into any other category. Given the intention to recruit participants between the ages of 5 and 80+, answering options were provided in emojis and written labels (Figure 6).

Figure 6: Experimental set-up (Pavlovia)



Note. **The background of the experiment created with OpenSesame was black. All other aspects (i.e. emojis, labels, and question) were identical.**

During the experimental stage, participants were presented with 4 practice trials (2 female and 2 male voice stimuli) to familiarise themselves with the experimental set-up and make adjustments to the volume if necessary. They were then asked whether they wanted to continue to the experimental block of 35 voice stimuli (speaker sex was allocated randomly). At the end of the experiment, participants saw a "correct score" which was predominantly intended as an immediate outcome for the children. Participants were verbally debriefed and received a physical copy of a form with the main aims and contact details of the experimenter. Online participants were redirected to a website hosting the pdf version of the same document, which could be downloaded for safekeep.

3.2.6 Deviations from Pre-registration

The three-way interaction of the random slope by-item structure specified during pre-registration was overly complex and resulted in the model not computing (see model (1) below). We therefore modified the random by-item term (see model (2) below), however, after analysing the PCA of random-effects variance-covariance estimates (rePCA in the lme4 package; Bates et al., 2015), the listener age by listener sex interaction was still overfitting and was therefore removed, resulting in the final (otherwise maximal) model (see model (3) below). These adjustments were made for models including listener age as a

continuous variable as well as listener age groups. To avoid convergence issues, the optimizer optimx (Nash, 2014; Nash & Varadhan, 2011) was used.

(1) Correct Response ~ Listener Age * Emotion Category * Listener Sex * Speaker Sex + (1 + Emotion Category | Participant ID) + (1 + Emotion Category * Listener Age * Listener Sex | Voice ID)

(2) Correct Response ~ Listener Age * Emotion Category * Listener Sex * Speaker Sex + (1 + Emotion Category | Participant ID) + (1 + Emotion Category + Listener Age * Listener Sex | Voice ID)

(3) Correct Response ~ Listener Age * Emotion Category * Listener Sex * Speaker Sex + (1 + Emotion Category | Participant ID) + (1 + Emotion Category + Listener Age + Listener Sex | Voice ID)

Furthermore, instead of running model comparisons between the full model and the model without the significant predictor, we opted to report the Wald-test statistics already displayed in the model output of the binomial mixed effects model. Likelihood ratio tests are usually preferred due to having slightly higher statistical power (Gudicha et al., 2017), however when sample sizes are large, Wald tests and likelihood ratio tests produce similar results (Winter, 2019).

3.2.7 Data analysis plan and preparation

Data were analysed using R (Version 4.0.4; R Core Team, 2020) and RStudio (Version 1.1.463; RStudio Team, 2016) with packages car (Version 3.0.10; Fox & Weisberg, 2019), tidyverse (Version 1.3.1; Wickham et al., 2019), e1071 (Version 1.7.6; Meyer et al., 2021), lme4 (Version 1.1.26; Bates et al., 2015), optimx (Version 2020.4.2; Nash, 2014; Nash & Varadhan, 2011), broom.mixed (Version 0.2.6; Bolker & Robinson, 2020), and emmeans (Version 1.5.5.1; Lenth, 2021).

To determine “Correct Response”, we used the speaker’s intended emotion category as baseline. Responses matching with the intended emotion category were scored 1, all other emotion categories and the “I don’t know” option were scored 0. For the confusion matrices, we computed mean accuracy per emotion category, listener sex, speaker sex, and listener age group from those correct

responses. Unbiased hit rates (H_u; Wagner, 1993) and chance-corrected recognition rates (CCR; Lassalle et al., 2019) were calculated for each listener age group and emotion category. Computing both measures was done to allow for comparisons with previous literature reporting either option. The following formulae were used:

$$H_u = \frac{number\ of\ correct\ responses}{\substack{number\ of\ times\ the \\ emotion\ category\ was\ presented}} * \frac{number\ of\ correct\ responses}{\substack{number\ of\ times\ that \\ emotion\ response\ was\ made}}$$

$$CCR = \frac{\frac{Accuracy\ in\ Percent}{100} - \frac{number\ of\ correct\ choice}{number\ of\ all\ choices}}{\frac{number\ of\ incorrect\ choices}{number\ of\ all\ choices}}$$

Two different binomial mixed-effects models with by-participant and by-item random intercepts and slopes were computed: one treating listener age as a continuous variable (age 5-39 years); the other dividing participants into age groups of children (5-10 years), adolescents (11-19 years), and young adults (20-39 years). For the binomial mixed-effects models, data were not averaged. Listener age as a continuous variable was centred, and mean-centred deviation coding was used for listener sex, speaker sex, emotion category (reference category: neutral), and listener age group (reference category: young adults).

Due to a programming error, the background colour of the experiment differed between testing locations (i.e. black in GSC, and grey on Pavlovia). Since participants recruited via Pavlovia were males between the ages of 18 and 39, we computed a binomial mixed effects model with by-participant and by-item random intercepts and slopes on Location (GSC, Pavlovia) for this specific age range and listener sex (see model (4) below).

(4) Correct Response ~ Listener Age * Emotion Category * Speaker Sex *
Location + (1 + Emotion Category | Participant ID) +
(1 + Emotion Category + Listener Age | Voice ID)

The probability to give a correct response did not depend on Location. The main effect of Location (Wald $X^2(1) = 0.048$, $p = .827$), and interactions of Location

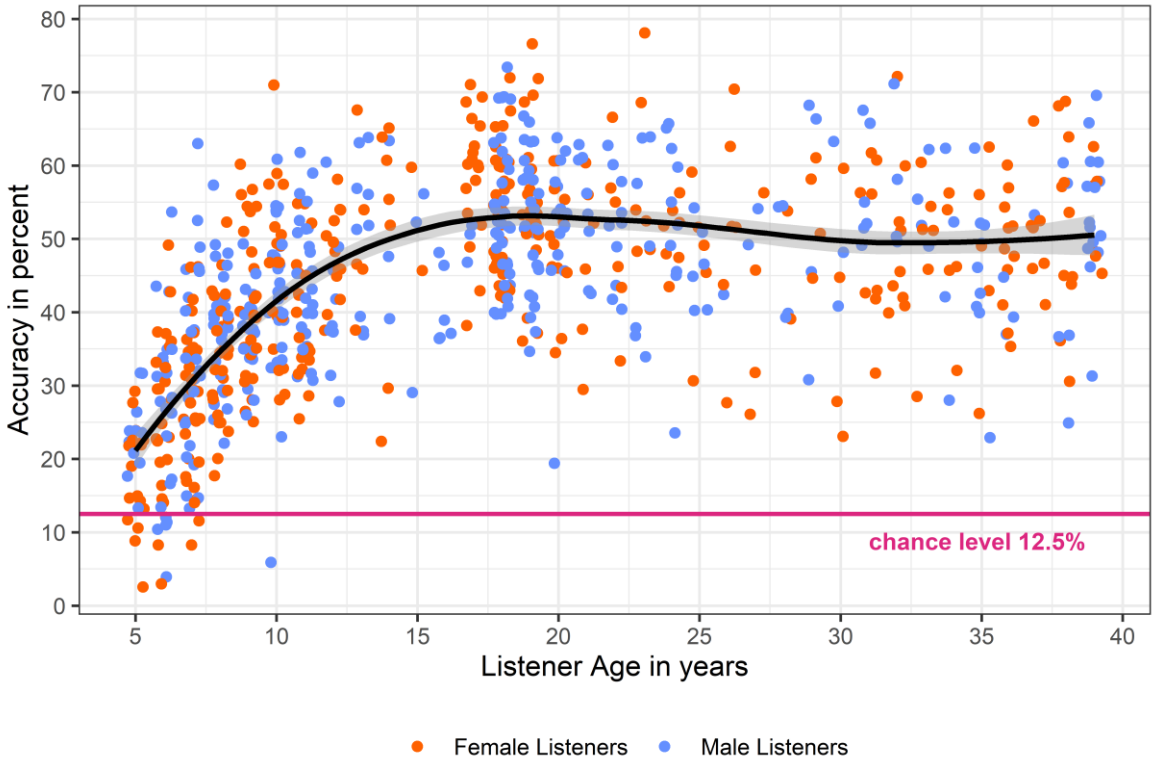
with Listener Age, Emotion Category and/or Speaker Sex were non-significant (all $p > .05$).

3.3 Results

3.3.1 Descriptive statistics

All three age groups performed well above chance levels (12.5% in an 8-AFC): Children displayed an overall average accuracy of 33.3% (n = 284, SD = 12.3%, median = 34.3%, skewness = -0.112, kurtosis = -0.013), adolescents 50.3% (n = 246, SD = 10.5%, median = 51.4%, skewness = -.0452, kurtosis = -0.044) and young adults 50.0% (n = 252, SD = 10.4%, median = 51.4%, skewness = -0.112, kurtosis = -0.373). Plotting accuracy in relation to age as a continuous variable (Figure 7), we can see recognition accuracy to increase with increasing age with a steeper incline between childhood and adolescence, and a less steep one during adolescence. Overall accuracy seems to peak in late adolescence and then plateaus during adulthood. Whilst the average of the youngest children is at recognition accuracy well above chance levels, there are still individual children who perform below (see data points below the magenta line in Figure 7).

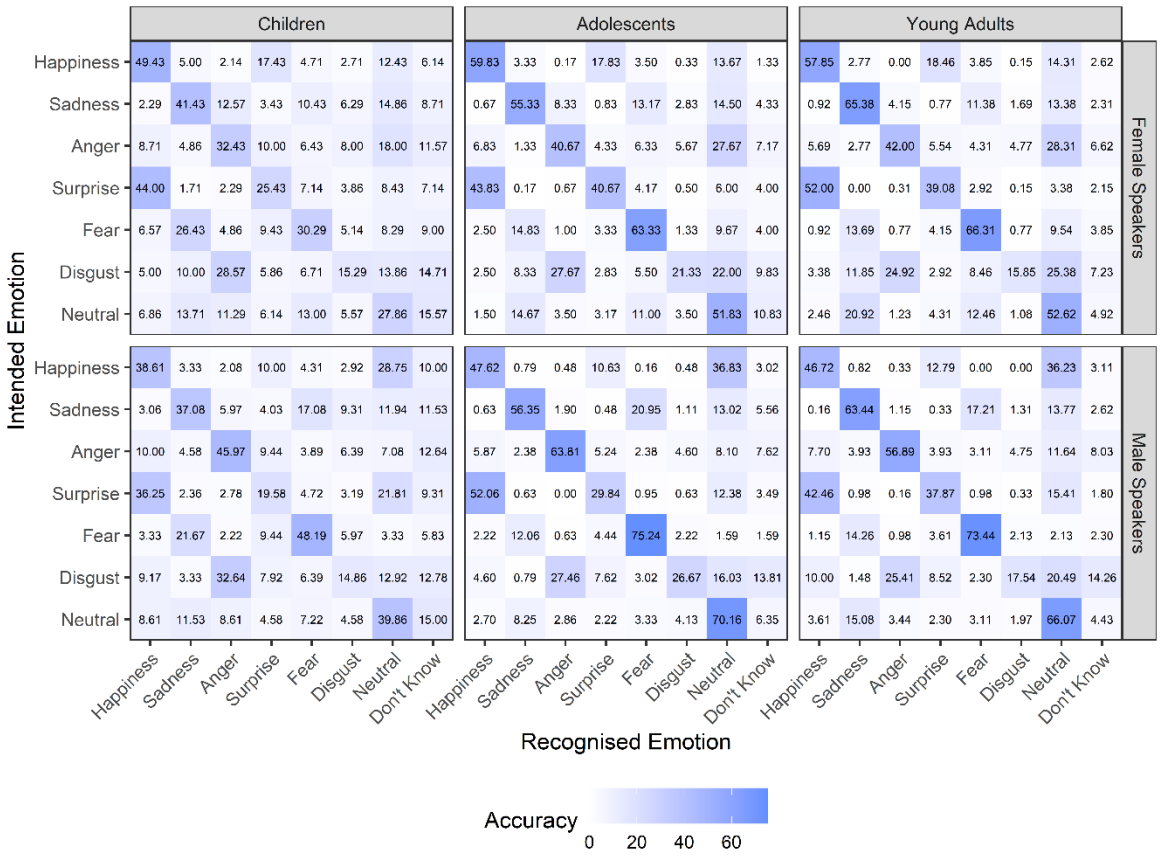
Figure 7: Emotion recognition accuracy by continuous listener age



Note. **Each point represents a listeners' accuracy score. A trend line (black) and a horizontal line indicating chance level (magenta) were added.**

Breaking down accuracy by emotion category, age group, and speaker sex, the confusion matrix (Figure 8) shows emotion categories to be well recognised in children, adolescents, and young adults for stimuli uttered by both male and female vocal stimuli (i.e. correct responses on the diagonal), yet, accuracy improves from childhood to adolescence and young adulthood (i.e. diagonal values in Figure 8 increase). Exceptions are surprise and disgust. Across all age groups and speaker sexes, surprise gets misclassified as happiness, and disgust as anger. Additionally, for adolescents and young adults, disgust seems to get confused with neutral. Children's perception of fear, when portrayed by a female but not a male speaker, appears confused with sadness. In general, fear and anger appear to be recognised better when spoken by a male, whereas happiness uttered by female speakers appears to obtain higher recognition rates. For comparisons with previous literature, we added unbiased hit rates, chance-corrected recognition rates, and additional correlation matrices (i.e. split by listener sex, and a matrix of differences) in Supplementary material 2.

Figure 8: Confusion matrix of emotion recognition accuracy, separated by age group and voice sex



***Note*. Children (5-10 years), Adolescents (11-19 years), Young Adults (20-39 years).**

3.3.2 Analysis of age as a continuous variable

3.3.2.1 Binomial mixed effects model

For age as a continuous variable (henceforth referred to as “continuous model”), the binomial mixed-effects model revealed a main effect of age ($X^2(2, N = 782) = 71.05$, $p < .001$), a main effect of emotion ($X^2(6, N = 782) = 341.98$, $p < .001$), and two-way interactions of listener age and emotion ($X^2(12, N = 782) = 147.29$, $p < .001$), listener sex and emotion ($X^2(6, N = 782) = 32.92$, $p < .001$), and speaker sex and emotion ($X^2(6, N = 782) = 89.01$, $p < .001$). There were no further significant main effects or interactions.

3.3.2.2 Hypothesis H1a: main effect of age as a continuous variable

H1a: For age as a continuous variable, we expect emotion recognition ability for each emotion category to improve with increasing age.

Analysis of the continuous model returned a small, yet significant main effect of age (logit coefficient = 0.034, SE = 0.004, z = 8.429, $p < .001$). That means that the model-based predicted probability of perceiving emotion correctly (i.e. matching the speaker's intended emotion) is .452 for a child at age 5, and .725 for an adult at 39 years of age. Hypothesis 1a is therefore supported; recognition ability increases with increasing age.

3.3.2.3 Hypothesis H2: interaction of listener age and emotion

H2: There will be a significant two-way interaction between listener age and emotion category.

The continuous model returned a significant interaction between listener age and emotion. From Figure 9 and Table 7 it can be seen that the model-based predicted probability of perceiving emotion correctly increases with an increase in age for all emotion categories bar disgust. However, the trends of fear, sadness and neutral are significantly steeper than for happiness, anger, and surprise respectively (apart from surprise and neutral, $p = .091$). The trends within the two groupings did not differ significantly between emotion categories. Disgust recognition decreased slightly with increasing age which is

significantly different from the trends of all other emotion categories (all $p < .001$). For an overview for exact statistics, see Supplementary material 3. Hypothesis 2 is therefore supported for the continuous model.

Figure 9: Linear age trends for model-based predicted probability of correct response, separately for each emotion category

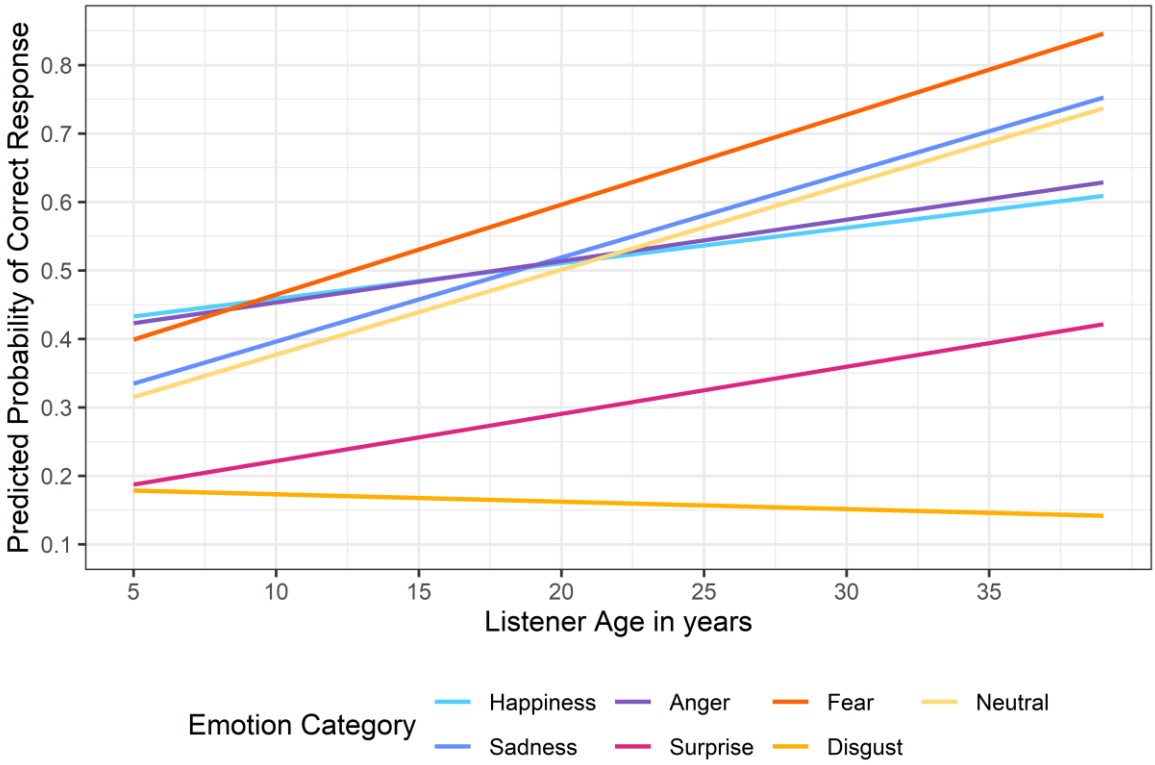


Table 7: Age trends estimates for model-based predicted probability of correct response, separately for each emotion category

Emotion	Age Trend	Standard Error	Asymp. LCL	Asymp. UCL
Happiness	0.021	0.006	0.009	0.033
Sadness	0.053	0.005	0.042	0.064
Anger	0.025	0.006	0.013	0.036
Surprise	0.034	0.006	0.022	0.045
Fear	0.062	0.006	0.050	0.074
Disgust	-0.008	0.007	-0.021	0.005
Neutral	0.053	0.006	0.041	0.065

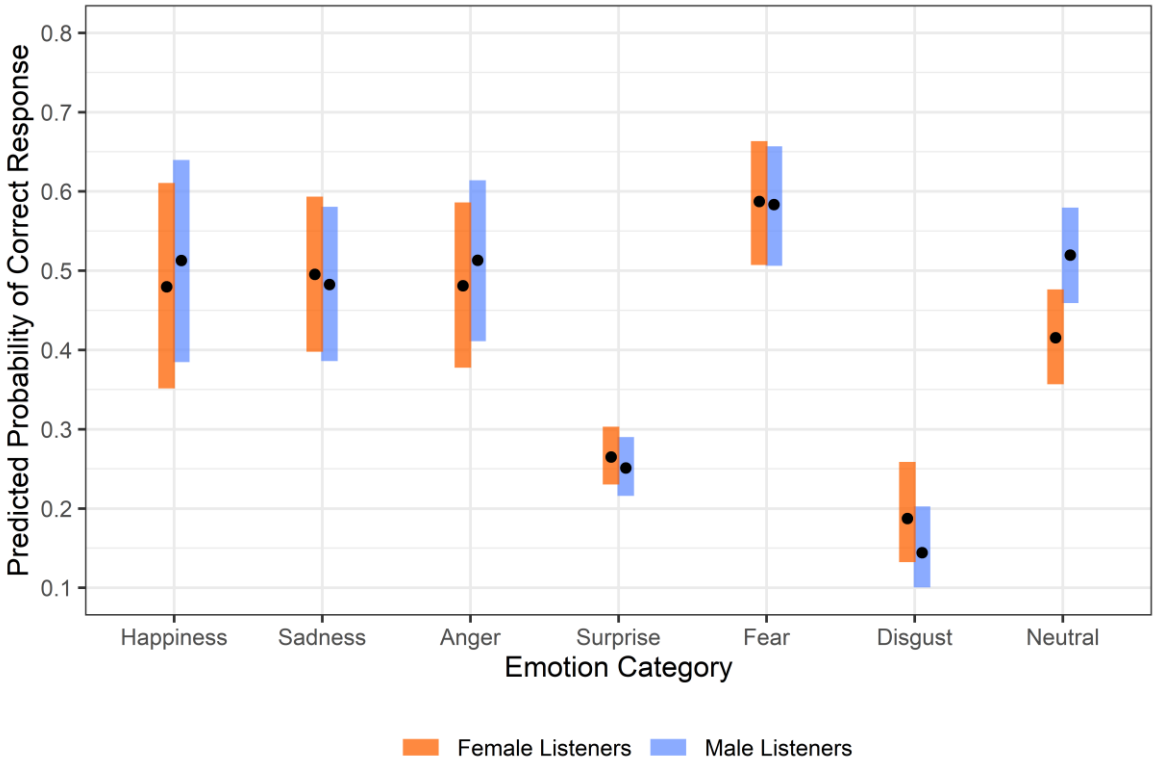
Note. **Asymp. LCL = asymptotic lower confidence interval. Asymp. UCL = asymptotic upper confidence interval. Df = inf for asymptotic test.**

3.3.2.4 Hypothesis H3: main effect of listener sex

H3: Recognition accuracy will differ between female and male listeners. However, given the contradictory literature, we will refrain from a specific directionality for this hypothesis.

There was no statistically significant effect of listener sex in the continuous model. There was, however, a significant interaction of listener sex and emotion category. Female listeners had a higher probability to give a correct response on disgusted (estimate = 0.043, SE = 0.016, z = 2.798, p = .035), whereas male listeners were predicted to be better decoders for neutral vocalisations (estimate = 0.104, SE = 0.025, z = 4.166, p < .001). There were no significant gender differences for any other emotion category (all sidak-corrected p-values > .05; see Figure 10). Since there was no significant main effect of listener sex, hypothesis 3 is rejected for the continuous model.

Figure 10: Model-based predicted probability of correct response for female and male listeners, separately for each emotion category



Note. **Coloured bars represent 95% asymptotic confidence intervals.**

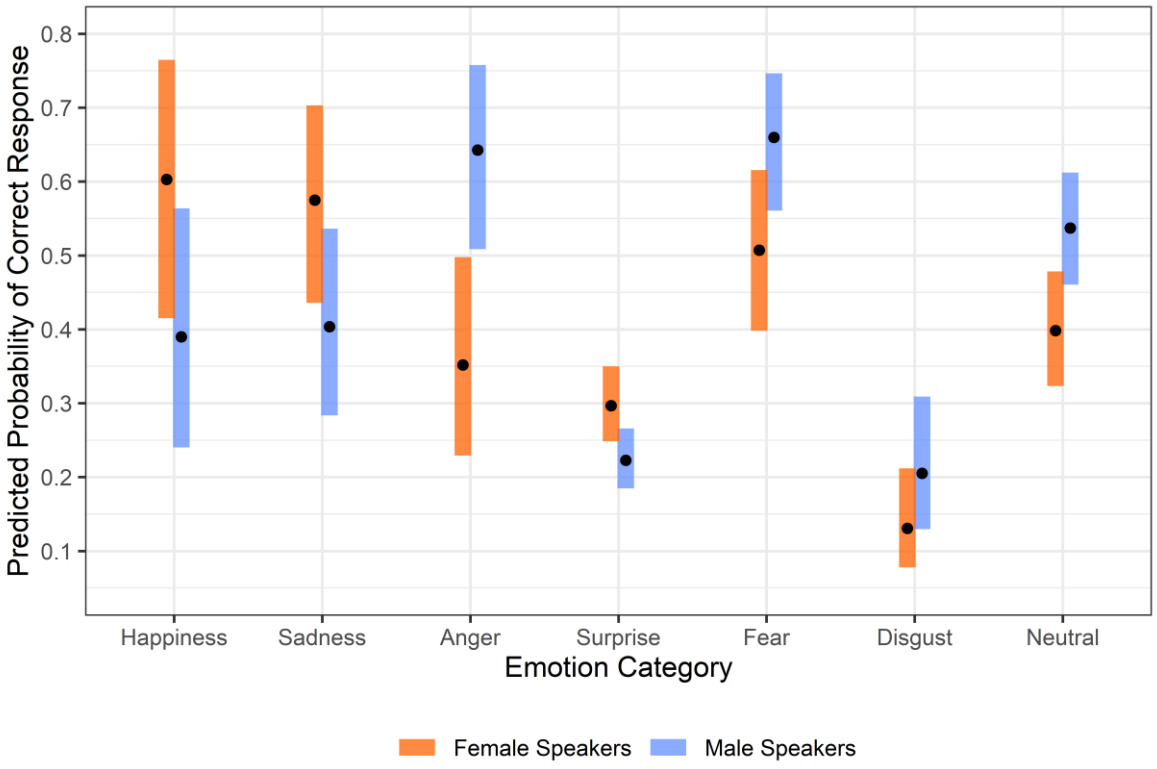
3.3.2.5 Hypothesis H4: main effect of speaker sex

H4: Recognition accuracy will differ between female and male speakers. However, given the contradictory literature, we will refrain from a specific directionality for this hypothesis.

Similar to the analysis of listener sex, there was no significant main effect of speaker sex. There was, however, a significant interaction between speaker sex and emotion.

Figure 11 shows that happy, sad, and surprised stimuli had a higher model-based predicted probability of being identified correctly when encoded by a female speaker whereas anger, fear, disgust, and neutral were recognised better when uttered by a male speaker. However, contrast comparison after sidak-corrections revealed that only the difference for anger remained significant. All other comparisons revealed no significant differences between female and male speakers (see Table 8). Therefore, hypothesis 4 of a significant main effect of speaker sex is rejected for the continuous model.

Figure 11: Model-based predicted probability of correct response for female and male speakers, separately for each emotion category



Note. **Coloured bars represent 95% asymptotic confidence intervals.**

Table 8: Contrast comparisons for model-based predicted probability between male and female speakers; separately for each emotion category

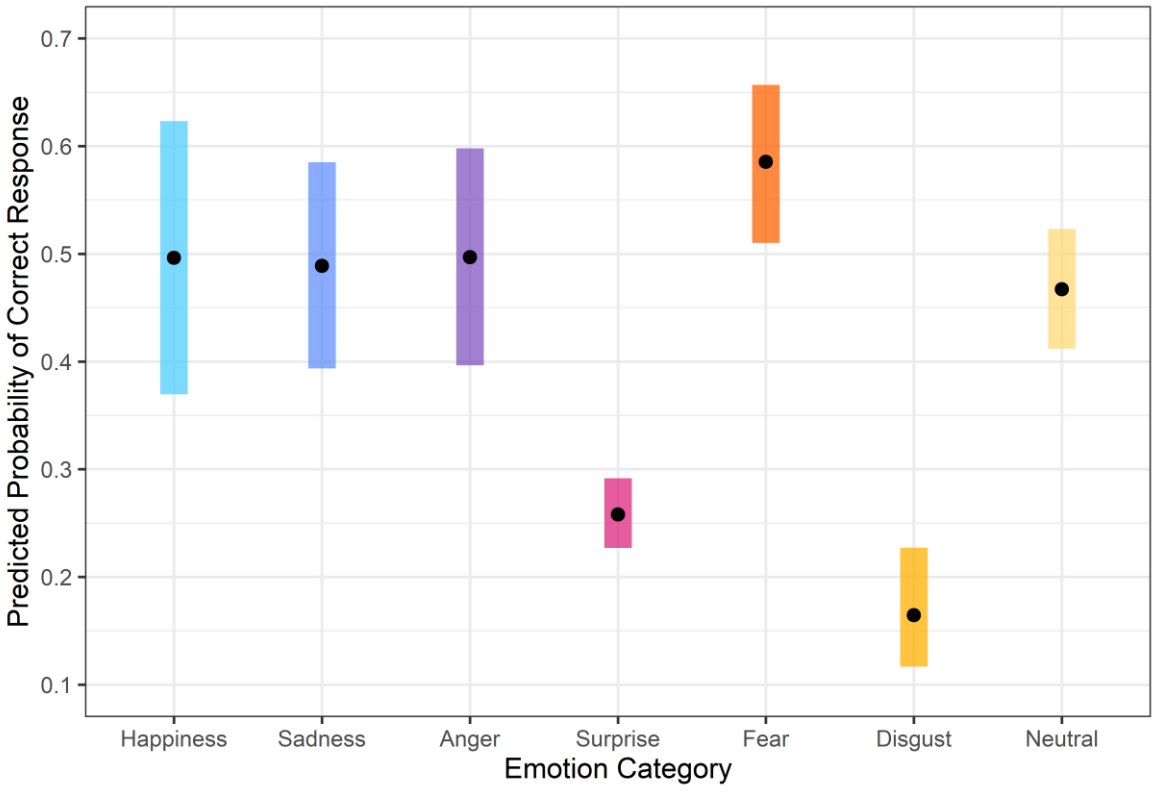
Emotion	Estimate	Standard Error	Z ratio	P value
Happiness	-0.213	0.126	-1.687	.489
Sadness	-0.172	0.096	-1.786	.416
Anger	0.291	0.095	3.052	.016
Surprise	-0.074	0.033	-2.228	.168
Fear	0.153	0.074	2.062	.244
Disgust	0.074	0.057	1.308	.773
Neutral	0.139	0.056	2.492	.086

***Note*. All comparisons are sidak-corrected. The estimate is the difference of predicted probability to respond correctly between male and female speakers. Negative values show a male advantage. Df = inf for asymptotic test.**

3.3.2.6 Additional findings not hypothesised

Additionally, the model returned a significant main effect of emotion. Tukey-corrected contrast comparisons showed surprise and disgust were significantly different from all other emotion categories and from one another (see Figure 12, and Supplementary material 4 for contrast tables).

Figure 12: Model-based predicted probability of correct response per emotion category



3.3.3 Analysis of age groups

3.3.3.1 Binomial mixed effects model

Similar to the continuous model, the binomial mixed-effects model treating age as grouping variable (henceforth called “age group model”) revealed a main effect of listener age group ($X^2(2, N = 782) = 155.39$, $p < .001$), a main effect of emotion ($X^2(6, N = 782) = 321.30$, $p < .001$), and two-way interactions of listener age group and emotion ($X^2(12, N = 782) = 164.22$, $p < .001$), listener sex and emotion ($X^2(6, N = 782) = 29.90$, $p < .001$), and speaker sex and emotion ($X^2(6, N = 782) = 100.98$, $p < .001$). In contrast to the continuous model, this model returned an additional two-way interaction of listener age group and listener sex ($X^2(2, N = 782) = 7.21$, $p = .027$), and a three-way interaction of listener age group, speaker sex and emotion ($X^2(12, N = 782) = 26.86$, $p = .008$). There were no further significant main effects or interactions.

3.3.3.2 Hypothesis H1b: main effect of age group

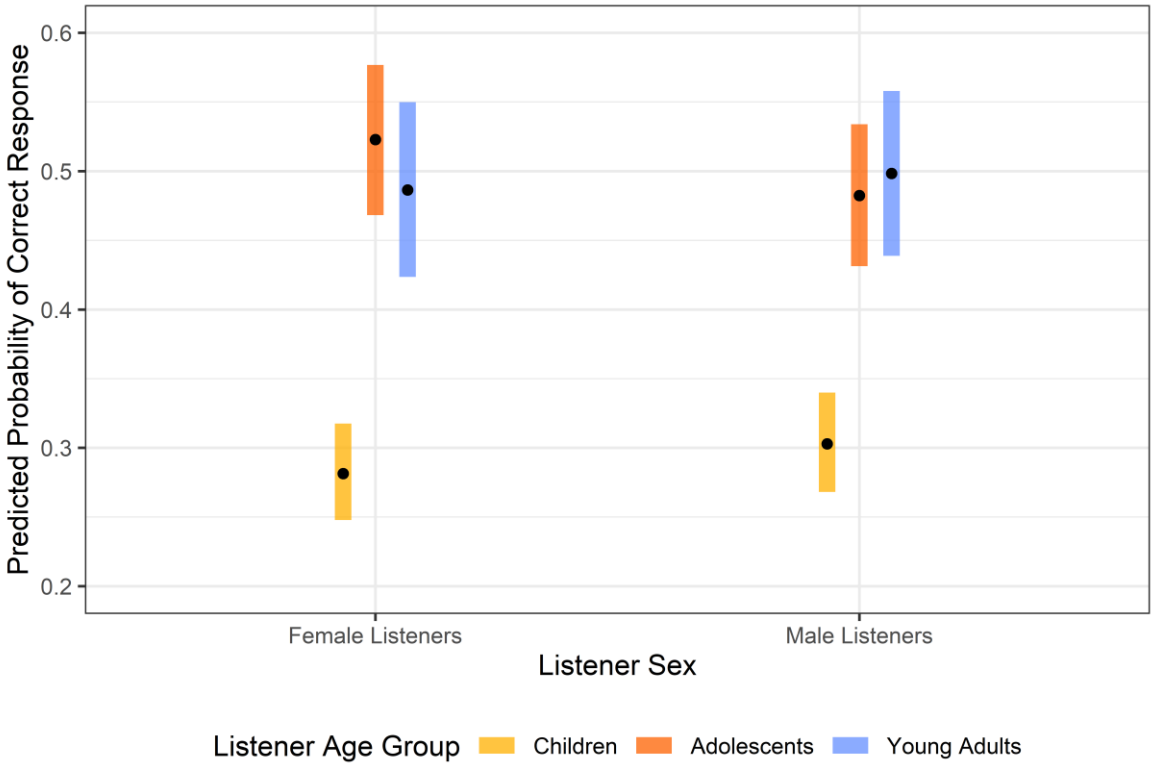
H1b: Children will be significantly lower in emotion recognition ability compared to adolescents and young adults. Differences between adolescents and young adults will be smaller.

The age group model revealed a main effect of age group which was investigated further. Tukey-corrected contrast comparison showed that adolescents were 2.45 ($p < .001$) and adults 2.35 ($p < .001$) times more likely to give a correct response compared to children. There was no significant difference between adolescents and adults (Odds Ratio = 1.04, $p = .801$).

Contrastingly to the continuous model, the analysis also returned a two-way interaction of listener age group and listener sex. We computed post-hoc comparisons between female and male listeners at each level of listener age group as well as comparisons between children, adolescents, and young adults at each level of listener sex. Within each age group, there were no significant differences between female and male listeners’ accuracy (all sidak-corrected p-values > .05). However, comparisons between the three levels of listener age group revealed that children had a significantly lower probability of providing a correct response compared to adolescents and young adults (see Figure 13, and

Table 9). Emotion recognition ability did not differ significantly between adolescent and adult listeners. These results held true for both female and male listeners showing that the interaction was driven by the age group effect.

Figure 13: Model-based predicted probability of correct response for each listener age group, separately for female and male listeners



Note. **Coloured bars represent 95% asymptotic confidence intervals.**

Table 9: Differences in model-based predicted probability of correct response for listener age group contrasts, separately for female and male listeners

Listener Age Group Contrast	Estimate	Standard Error	Z ratio	P value
		Female Listeners		
Children – Adolescents	-0.241	0.022	-11.209	<.001
Children – Young Adults	-0.205	0.026	-7.764	<.001
Adolescents – Young Adults	0.036	0.021	1.753	.526
		Male Listeners		
Children – Adolescents	-0.180	0.022	-8.362	<.001
Children – Young Adults	-0.195	0.027	-7.364	<.001
Adolescents – Young Adults	-0.016	0.021	-0.757	.995

Note. **All comparisons are sidak-corrected. Children (5-10 years); Adolescents (11-19 years); Young Adults (20-39 years). The estimate is the difference of predicted probability to respond correctly between the compared listener age groups. Df = inf for asymptotic test.**

Hypothesis 1b is therefore partially supported; children are significantly lower in emotion recognition accuracy compared to adolescents and young adults. However, adolescents and young adults did not differ significantly in emotion recognition accuracy.

3.3.3.3 Hypothesis H2: interaction of listener age and emotion

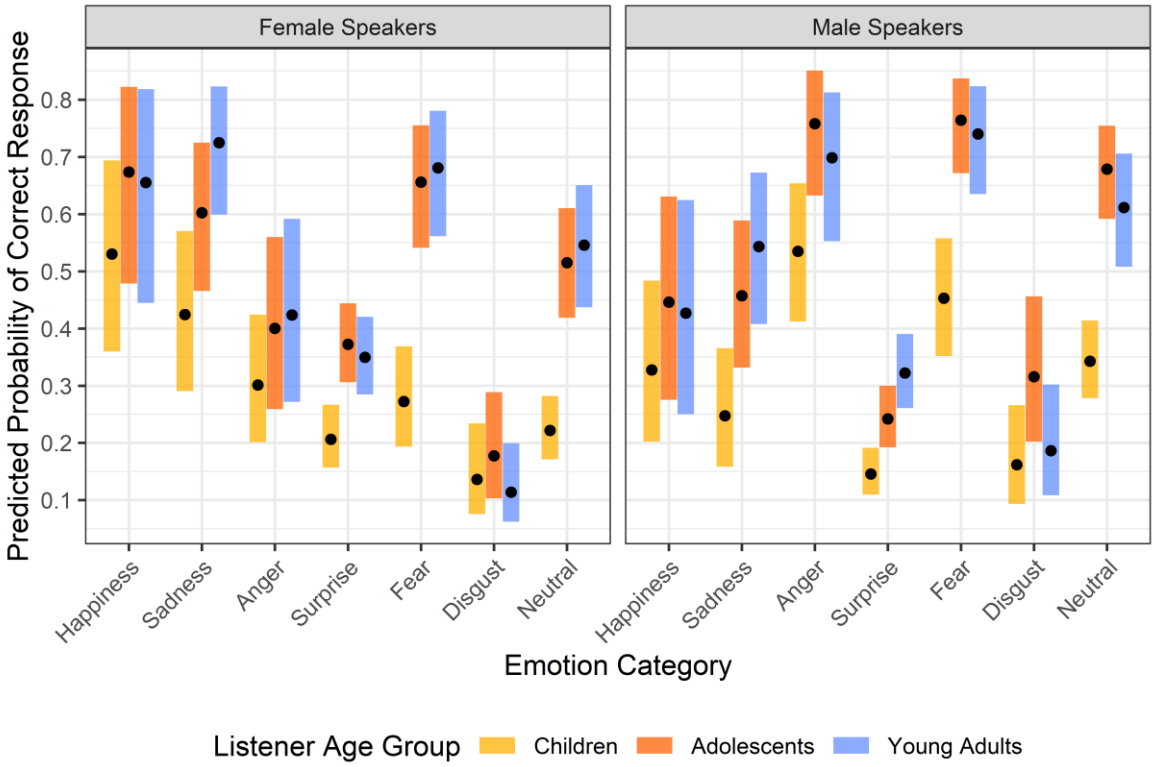
H2: There will be a significant two-way interaction between listener age and emotion category.

The age group model returned a significant interaction between age group and emotion category but contrastingly to the continuous model, this interaction was further quantified by speaker sex. To break down the three-way interaction, simple contrasts were used for age group at each level of emotion category and speaker sex. All contrast comparisons were sidak-corrected to account for multiple comparisons.

For happiness, children differed significantly from adolescents when vocal stimuli were produced by female but not male speakers. There were no significant differences between children and young adults or adolescents and young adults for either male or female speakers. Perception accuracy of sad utterances spoken by female speakers increased significantly between childhood and adolescence, childhood and adulthood, and adolescence and adulthood. Male speaker's sad expressions were better recognised by adolescents and young adults in comparison to children, however, there was no significant change between adolescents' and young adults' perception. There were no significant differences between the three age groups in the perception of anger or disgust from female speakers. The accuracy pattern of angry expressions mirrored the ones for sadness for the male speakers. Adolescents and young adults had higher recognition accuracy compared to children, however, they did not differ significantly from one another. Contrastingly, for disgusted expressions from male speakers, adolescents differed significantly from children and adults, but children and adults performed on par. For surprise, fear, and neutral, results showed significant differences between children and adolescents as well as children and young adults but not between adolescents and young adults. This held true for stimuli encoded by female and male speakers (Figure 14 and Table

10; additional contrast comparisons were added in Supplementary material 5). Hypothesis 2 is therefore also supported for the model including listener age groups.

Figure 14: Model-based predicted probability of correct response of the interaction of age group by emotion category, separated by speaker sex



Note. **Coloured bars represent 95% asymptotic confidence intervals.**

Table 10: Differences in model-based predicted probability of correct response for listener age group contrasts for each emotion category, separately for female and male speakers

Age Group Contrast	Female Speakers				Male Speakers			
	Est	SE	Z ratio	P value	Est	SE	Z ratio	P value
				Happiness				
CH - AD	-0.143	0.040	-3.557	.016	-0.118	0.044	-2.694	.258
CH – YA	-0.125	0.045	-2.757	.218	-0.099	0.050	-1.991	.865
AD - YA	0.018	0.040	0.453	1.000	0.019	0.042	0.458	1.000
				Sadness				
CH - AD	-0.178	0.039	-4.551	<.001	-0.210	0.035	-5.991	<.001
CH – YA	-0.301	0.044	-6.826	<.001	-0.295	0.041	-7.121	<.001
AD - YA	-0.123	0.035	-3.472	.021	-0.086	0.037	-2.327	.571
				Anger				
CH - AD	-0.099	0.043	-2.317	.581	-0.223	0.035	-6.426	<.001
CH – YA	-0.123	0.049	-2.489	.419	-0.164	0.041	-3.975	.003
AD - YA	-0.023	0.041	-0.578	1.000	0.060	0.035	1.716	.977
				Surprise				
CH - AD	-0.166	0.035	-4.732	<.001	-0.096	0.028	-3.461	.022
CH – YA	-0.143	0.039	-3.689	.009	-0.176	0.035	-5.028	<.001
AD - YA	0.023	0.038	0.597	1.000	-0.080	0.034	-2.324	.575
				Fear				
CH - AD	-0.384	0.039	-9.958	<.001	-0.311	0.039	-8.072	<.001
CH – YA	-0.409	0.042	-9.623	<.001	-0.287	0.043	-6.690	<.001
AD - YA	-0.025	0.037	-0.665	1.000	0.024	0.034	0.713	1.000
				Disgust				
CH - AD	-0.041	0.026	-1.579	.994	-0.154	0.039	-3.986	.003
CH – YA	0.022	0.025	0.909	1.000	-0.025	0.032	-0.783	1.000
AD - YA	0.064	0.026	2.443	.460	0.129	0.037	3.485	.020
				Neutral				
CH - AD	-0.293	0.044	-6.689	<.001	-0.336	0.040	-8.314	<.001
CH – YA	-0.324	0.049	-6.557	<.001	-0.268	0.048	-5.586	<.001
AD - YA	-0.031	0.044	-0.691	1.000	0.067	0.043	1.572	.994

***Note*. All comparisons are sidak-corrected. CH – Children (5-10 years); AD = Adolescents (11-19 years); YA = Young Adults (20-39 years). Est = Estimate is the difference of predicted probability to respond correctly between the compared listener age groups. SE = Standard Error. Df = inf for asymptotic test.**

3.3.3.4 Hypothesis H3: main effect of listener sex

H3: Recognition accuracy will differ between female and male listeners. However, given the contradictory literature, we will refrain from a specific directionality for this hypothesis.

Similar to the continuous model, this model did not return a main effect of listener sex, but a listener age group by listener sex interaction. Hypothesis 3 is therefore not supported for the age group model. For details, see chapter 3.3.2.4. Whilst values may differ slightly, outcomes, directionality, and interpretation remain the same.

3.3.3.5 Hypothesis H4: main effect of speaker sex

H4: Recognition accuracy will differ between female and male speakers. However, given the contradictory literature, we will refrain from a specific directionality for this hypothesis.

Equivalent to the continuous model, there was no significant main effect of speaker sex, but a two-way interactions of speaker sex by emotion. Again, Hypothesis 4 is not supported for the age group model. Since the age group model and the continuous model returned similar results, see chapter 3.3.2.5 for results and interpretation (numeric values may differ slightly, but directionality and interpretation remain the same).

3.3.3.6 Additional findings not hypothesised

The age group model also returned a significant main effect of emotion. Similar to the continuous model, tukey-corrected contrast comparisons showed surprise and disgust were significantly different from all other emotion categories. In contrast though, surprise and disgust did not differ significantly from one another (estimate = 0.090, SE = 0.033, z = 2.742, p = .088).

3.4 Discussion

This study aimed to investigate the developmental course of perceived vocal emotion recognition between childhood and early adulthood by using the socially-relevant word “hello”. We found a main effect of listener age when

using a binomial mixed-effect model treating age as a continuous variable that was also present in the model dividing participants into pre-determined age groups. We also found a significant interaction of listener age/ age group with emotion showing that different emotion categories mature at different rates. Contradictory to hypotheses 3 and 4, we did not detect any main effects of listener sex or speaker sex, however, both variables showed significant interactions with emotion. Taken together, these results suggest that vocal affect recognition matures with increasing age but varies for different emotion categories.

3.4.1 Overall accuracy and listener age effect

Overall, percentage of accuracy was lower in each emotion category for children compared to adolescents and young adults, yet all listener age groups performed at above-chance accuracy. This is in agreement with most research published (e.g. Chronaki et al., 2015; Nelson & Russell, 2011; Zupan, 2015; but see Aguert et al., 2013). The model-based predicted probability of correctly identifying vocal affect categories increased with an increase in age. Contrast comparisons from the listener age group model confirmed a significant difference between children and adolescents, as well as children and young adults, however not between adolescence and young adulthood. This shows that whilst children are able to recognise vocal emotion at better than chance levels, recognition accuracy improves between childhood and adolescence and not much thereafter.

The results are in line with previous papers that assume developmental maturation to occur between childhood and adolescence (e.g. Amorim et al., 2021; Grosbras et al., 2018; Morningstar, Ly, et al., 2018). Visualisation of accuracy in relation to continuous age suggested that adult-like ability is reached at around 16 years of age, however, since our model predictions were based on linear estimations and group comparisons, this cannot be determined precisely. Yet, our results contribute to the existing literature, showing a listener age effect in affect perception by using brief socially-relevant vocalisations.

We also need to highlight that overall recognition accuracy is slightly below rates usually reported in vocal emotion research. Our results suggested 33% recognition accuracy for children, and 50% for adolescents and young adults, whereas other studies (Amorim et al., 2021; Chronaki et al., 2018; Cortes et al., 2021; Demenescu et al., 2014; Grosbras et al., 2018; Zupan, 2015) report young children's emotion recognition accuracy at around 56% on average, and young adult listeners at around 73% or above. Yet, our findings are on par with recognition accuracy or chance-corrected recognition rates for adolescents and young adults reported elsewhere (Juslin & Laukka, 2001; Lassalle et al., 2019; Morningstar, Ly, et al., 2018). These differences in recognition accuracy may be explained by the variability in the data for separate emotion categories, stimulus choice, background noise of the testing environment, and/or task complexity.

Our results showed large variability between as well as within emotion categories which in turn could impact overall accuracy. Whilst fear and sadness were very well recognised across both speaker sexes and listener age groups comparable to previous literature (Gold et al., 2012; Sauter et al., 2013), disgusted and surprised vocalisations obtained the lowest accuracy scores. Recognition rate of happiness was slightly lower than fear and sadness, yet representative when more than one positive emotion category is presented in the testing paradigm (as observed in Belin et al., 2008; Cortes et al., 2021; Paulmann & Pell, 2010; Paulmann & Uskul, 2014). Confusion patterns of surprise as happiness, and disgust as anger are maintained across both speaker sexes and age groups, and are frequently reported in the literature (e.g. Juslin & Laukka, 2001; Pell et al., 2009). When investigating listener age as a continuous variable, steeper development was observed for emotion categories sadness, fear, and neutral, whereas happiness, anger, and surprise matured less steeply. Whilst surprise recognition improved with increasing age, disgust remained a poorly recognised emotion category across age. Still, the low performance of surprise and disgust could have skewed our results to having lower overall recognition accuracy.

Recognition rates may also vary due to stimulus choice. In this study, we employed speech stimuli of socially-relevant words. Hence, the overall

recognition rates reported here are closer in value to studies using speech stimuli than studies utilising non-verbal vocalisations. Indeed, research with affect bursts frequently obtains higher recognition rates compared to speech segments in adults (Hawk et al., 2009) and in children (Sauter et al., 2013). However, confusion matrices from an adult listener group with a variety of stimuli suggest this may not apply to all emotion categories equally (Lausen & Hammerschmidt, 2020). Specifically, disgust and surprise from affect bursts seem very well recognised (Lausen & Hammerschmidt, 2020; see also Belin et al., 2008; Lima et al., 2014; Sauter, Eisner, Calder & Scott, 2010), yet when encoded in speech, they appear the lowest accurately recognised emotion categories (Lausen & Hammerschmidt, 2020; see also Demenescu et al., 2014; Juslin & Laukka, 2001; Lambrecht et al., 2012, 2014; Paulmann & Uskul, 2014; Pell et al., 2009). Using word stimuli in this study could explain the low recognition rates of disgust and surprise. However, whilst recognition accuracy may be lower than in other studies, it is above chance-levels which still allows us to observe ageing effects.

Another explanation as to why some stimuli had low recognition rates may be sought in the background noise that listeners experienced within the Glasgow Science Centre compared to the pre-validating approach in a quiet office setting. However, humans have been able to detect vocal emotions even under noisy conditions (Liuni et al., 2020). This limitation would also not explain why some of the stimuli were recognised extremely well. Furthermore, the majority of trustworthiness ratings (Chapter 2) were gathered in the Glasgow Science Centre, and those did not seem to be impacted by the testing environment.

A final reason for the lower recognition rates in our study in comparison to other findings could be task complexity. Usually, when studying developmental trajectories of vocal emotion perception, researchers tend to use paradigms with subsets of two to four of the Ekman and Friesen (1971) canonical emotions (Chronaki et al., 2015; Morningstar, Ly, et al., 2018; Nelson & Russell, 2011; Quam & Swingley, 2012). It could be speculated that more affect categories to select from would result in higher cognitive load and subsequently lead to lower overall accuracy. We did not measure this explicitly, though, and suggest future research should investigate further.

3.4.2 No main effect of listener sex

Contradictory to hypothesis 3, we did not detect a main effect of listener sex. We did, however, find a listener sex by emotion interaction showing a very small advantage for females recognising disgust, and for males decoding neutral expressions. Happy, sad, angry, fearful, and surprised perceptions did not differ significantly between female and male listeners. This is in contrast to literature seeing no main effect of listener sex without any further interactions (e.g. Amorim et al., 2021; Sauter et al., 2013) but also to studies reporting an overall female advantage (e.g. Belin et al., 2008; Grosbras et al., 2018; Lausen & Schacht, 2018; Paulmann & Uskul, 2014). Comparably to our results, Sen et al. (2018) reported a significant listener sex by emotion interaction for their young adult listeners. Whilst our results are in line with their non-significant differences for anger and fear, Sen et al. (2018) showed young adult females to be more accurate decoders than their male counterparts for happy and neutral emotion which opposes our findings. One could argue the results differ due to our study including children and adolescents as well as young adults, however our age group by listener sex interaction showed females' and males' perception on par across all age groups. More importantly, when treating age as a continuous variable, the age by listener sex interaction was non-significant.

One possible factor to explain the absence of listener sex effects in this study may be sought in the repetition of stimuli. Studies presenting vocal representations only once, appear to report either overall listener sex differences (Grosbras et al., 2018; Lausen & Schacht, 2018; Paulmann & Uskul, 2014) or differences in specific emotion categories (Sen et al., 2018; the current study). Conversely, studies with paradigms allowing for repetitions find no main effect of listener sex (Amorim et al., 2021; Sauter et al., 2013). However, this requires further investigation with more stringent paradigms. Additionally, given that our listener sex differences were detected in an emotion category that was generally poorly recognised by both male and female listeners (i.e. disgust) and for neutral expressions, we should not over-interpret the significant interaction of listener age and emotion, but rather emphasize how similar females' and males' perception is for the remaining five emotion categories.

3.4.3 No main effect for speaker sex

Similar to listener sex, there was also no main effect of speaker sex, however, there was a significant interaction of speaker sex and emotion. Recognition rates for anger were significantly higher when encoded by male compared to female speakers. This difference remained significant after correction for multiple comparison were applied. This is contradictory to a reported overall female advantage for affect bursts (Belin et al., 2008; Lausen & Schacht, 2018) and positive/neutral nouns (Lausen & Schacht, 2018) but also to an overall male advantage for negative nouns (Lausen & Schacht, 2018). This could suggest that socially-relevant words are semantically dissimilar to positive, negative, and neutral nouns, but also to affect bursts.

Explanations may be sought in evolutionary frameworks in which anger is closely related to aggression. Reactions to anger for males are more likely to result in physical aggression or risk-taking (Deffenbacher et al., 1996; Fessler et al., 2004), whereas females tend to engage more in indirect expressions of anger such as gossiping, ignoring, stonewalling (Archer, 2004). Males tend to show aggressive behaviours for example in intra-sexual competition, maintaining respect, or social status (e.g. Buss, 1989; Fischer & Rodriguez Mosquera, 2001). One could speculate that if males are more aggressive, they are intrinsically more angry than females. However, self-report measures show males and females to match how frequently they feel angry (Archer, 2004; Fischer & Evers, 2010).

Our finding may therefore be explained better by learned social roles rather than innate difference in experiencing anger itself. Stereotypically, in a Western society, boys are encouraged to display anger outwardly from an early age to be seen as assertive, but should not express vulnerability, sadness, or anxiety. In contrast, girls are more encouraged to express positive emotions, but internalise negative emotion (Chaplin, 2014). Showing anger explicitly may have repercussions for females, such as being labelled as “hostile” (Schieman, 2010). These perceived societal gender roles may have influenced our speakers’ encoding ability subconsciously. Despite being instructed to portray anger in a theatrical way, interpretations may thus have differed for male and female

speakers, resulting in stimuli from male speakers to be expressed more overtly compared to female speakers.

3.4.4 Limitations and future outlook

The present study is not without limitations. Despite it being a large-scale study, participant ages within the adolescent group were not evenly distributed. Approximately, 33.33% of listeners were 11 to 13 years old, 58.13% between 17 and 19, with the remaining 8.54% between ages of 14 and 16 (see Supplementary material 1). This could have swayed results to make the adolescent group appear more adult-like in recognition ability and masked subtle differences. Analysing the data treating age as a continuous variable may have counteracted that discrepancy, however, based on linear model assumptions, we cannot determine with certainty when vocal emotion recognition skill matures to adult-like levels. Given that visualisations suggested a potential plateauing at around age 16, this age group should be investigated further with a larger sample size between the crucial ages of 14 and 16 to settle the ongoing debate in the literature as to the exact age of reaching adult-like recognition ability.

Secondly, whilst aiming to expand the field of developmental vocal emotion recognition by including socially-relevant stimuli, we need to draw attention to the difficulties both female and male speakers experienced producing recognisable stimuli for disgust. One may seek evolutionary or linguistic explanations: In an evolutionary sense, disgust is a universal psychological mechanism helping with pathogen avoidance including clues detecting food spoilage or infection in others (Cepon-Robins et al., 2021; Darwin, 1872). We could speculate that a natural response would be withdrawal from such an environment, whereas the word “hello” would perhaps be used in socially-welcoming situations indicative of an approaching behaviour. Linguistically, disgust would be represented with vowel sounds such as [ɯ, u, ʌ, ɜ] and fricative [x, ɸ, h] or bilabial nasal [m] consonants to avoid anything entering the mouth (Goddard, 2014). Contrastingly, the word “hello” is comprised of vowel sounds [ə or ɛ, and əʊ] not related to representations of disgust, and lateral approximant consonant sounds [l], making it difficult to encode the word “hello” in a disgusted manner. Yet, “hello” works well for the remaining 6 emotion

categories in this study and should not be discarded as a valid stimulus in future research.

Lastly, we did not control for stimulus intensity when selecting vocalisations for this study, since its main purpose was to investigate emotion recognition ability in relation to age. This may have impacted our results though as there is some research indicating perceptual differences between high- and low-intensity stimuli (Chronaki et al., 2015; Holz et al., 2021; Juslin & Laukka, 2001; Zupan, 2015). Frequently, highly intense stimuli are reported to be recognised better (e.g. Juslin & Laukka, 2001), however, a recent study (Holz et al., 2021) found that peak intensity is most ambiguous for recognising vocal emotion. Nevertheless, across age, there is still ambiguity how intensity connects to accuracy. One study (Zupan, 2015) found that only preschool children (ca. 5 years old) would perform significantly better with high compared to low intensity whereas another one (Chronaki et al., 2015) showed 10- to 11-year-old children were less accurate in emotion recognition compared to adults when listening to 50% and 75% morphs of emotional states (between emotion and neutral). However, since previous studies used affect bursts from female speakers only (Chronaki et al., 2015; Holz et al., 2021) or sentence stimuli with small listener sample sizes (Juslin & Laukka, 2001; Zupan, 2015), we cannot be entirely certain whether those interpretations are applicable to the socially-relevant word stimuli used in this study. The field would benefit from investigating the role of intensity in the developmental course of vocal emotion recognition from word stimuli.

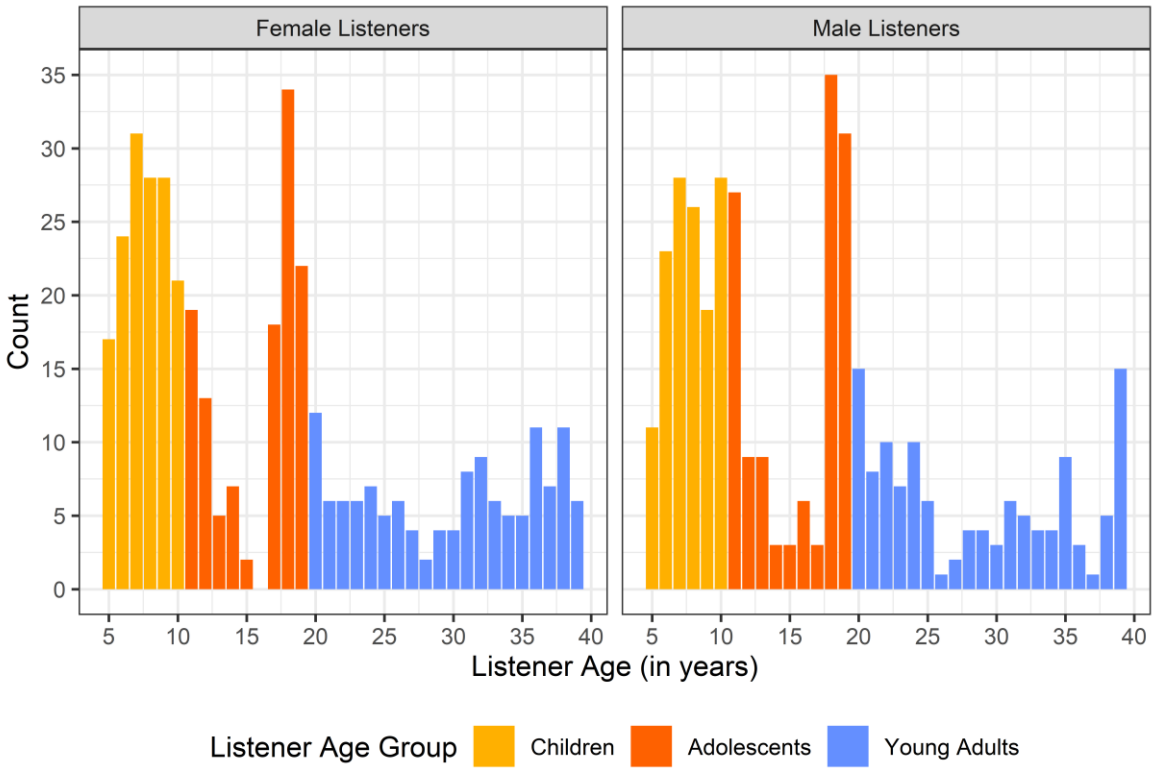
3.4.5 Conclusion

In summary, this study aimed to investigate the early developmental course of vocal emotion recognition abilities. We found a main effect of listener age (group): Despite children being better than chance-levels at correctly identifying emotion, recognition abilities improve significantly until adolescence, yet not much thereafter. We have also shown that the developmental progress depended on specific emotion categories. We did not detect any main effects of listener sex or speaker sex. Yet, we determined that male speakers encoded angry stimuli that had a higher probability of getting recognised, and that both male and female speakers struggled to produce recognisable stimuli for disgust.

Despite its limitations, this large-scale study enhances the field of vocal emotion recognition across the early lifespan by including brief socially-relevant stimuli.

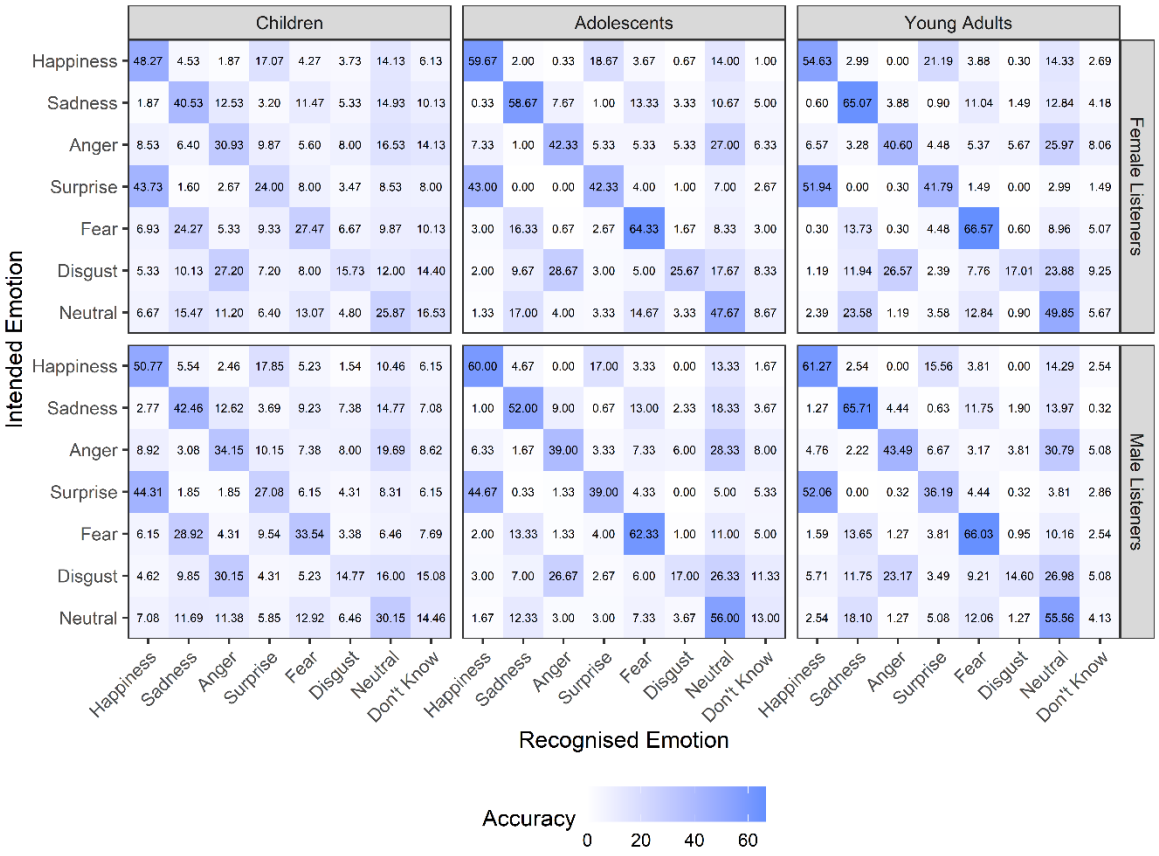
3.5 Supplementary material 1

Figure 15: Number of participants (continuous listener age)



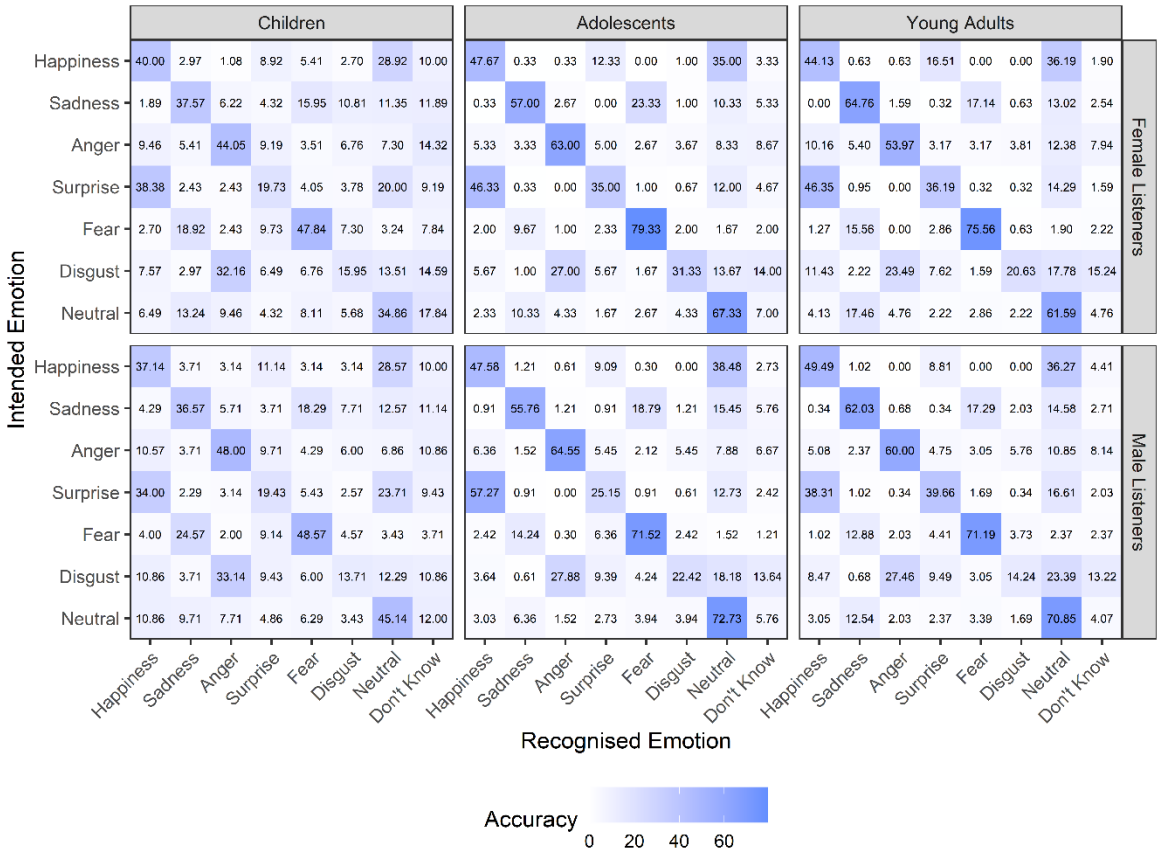
3.6 Supplementary material 2

Figure 16: Confusion matrix for female speakers



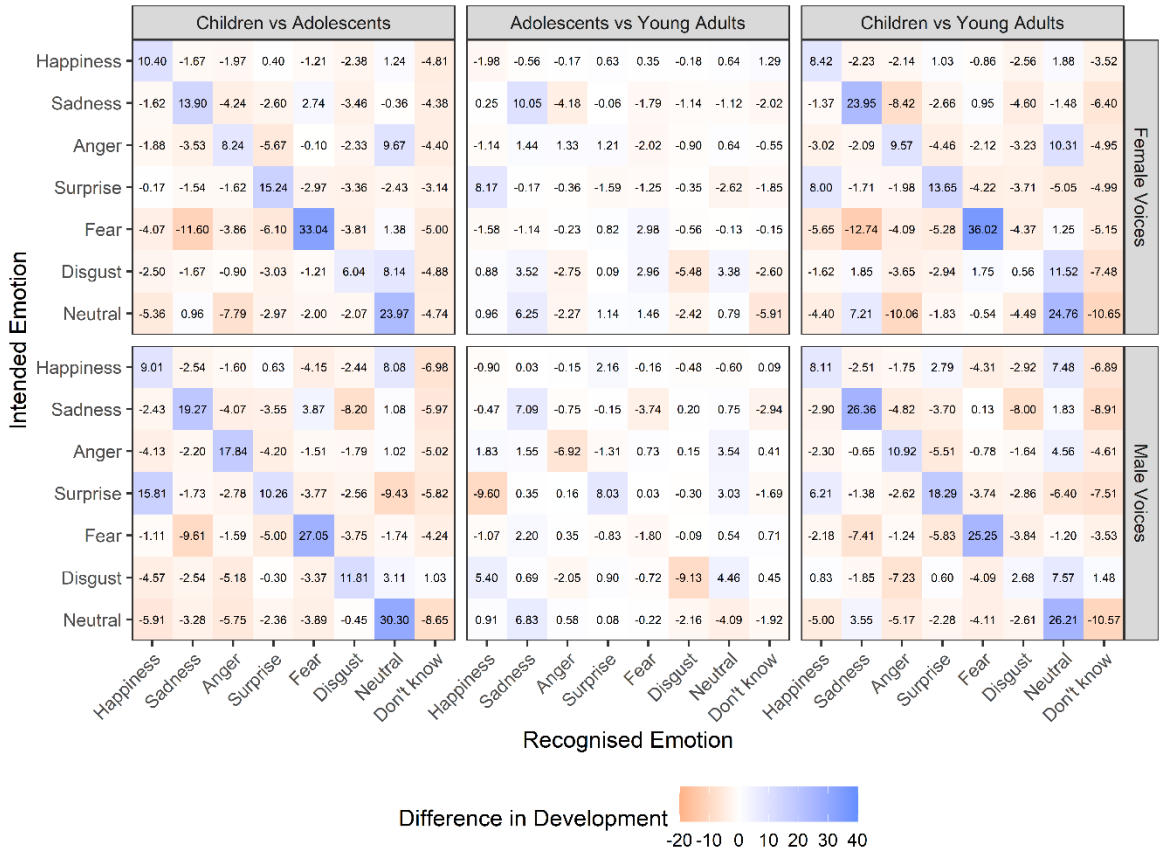
We can see here that the confusion patterns within each age group are very similar between male and female listeners for recognising emotion from female speakers.

Figure 17: Confusion matrix for male speakers



Comparably to Figure 16, confusion patterns within each age group are very similar for male and female listeners for recognising emotion from male speakers.

Figure 18: Difference matrix



In this confusion matrix, we can see the maturation of emotion categories between childhood and adolescence, adolescence and adulthood, and childhood and adulthood. Blue shades show “improvement”, whereas red shades show “decline” of recognition accuracy. Here, we see major improvements between childhood and adolescents with less so between adolescents and young adults. Mainly "correct recognition" (i.e. diagonal ratings) improves with age whereas recognition accuracy decreases in the patterns that were initially confused during childhood.

Table 11: Overall accuracy (ACC), chance-corrected recognition rates (CCR), and unbiased hit rates (H_u) of each emotion category for children, adolescents, and young adults, separately for female and male speakers

	Female Speakers				Male Speakers			
Emotion	ACC	SD_{ACC}	CCR	H_u	ACC	SD_{ACC}	CCR	H_u
				Children				
Happiness	49.43	26.11	42.21	19.89	38.61	27.24	29.84	13.67
Sadness	41.43	21.41	33.06	16.64	37.08	24.12	28.09	16.39
Anger	32.43	23.93	22.78	11.17	45.97	26.98	38.25	21.08
Surprise	25.43	22.48	14.78	8.32	19.58	22.74	8.09	5.90
Fear	30.29	25.89	20.33	11.65	48.19	28.72	40.79	25.30
Disgust	15.29	18.05	3.19	4.99	14.86	17.82	2.70	4.68
Neutral	27.86	26.79	17.55	7.48	39.86	30.79	31.27	12.64
				Adolescents				
Happiness	59.83	21.30	54.09	30.43	47.62	21.89	40.14	19.60
Sadness	55.33	24.01	48.95	31.24	56.35	23.38	50.11	39.07
Anger	40.67	24.04	32.19	20.17	63.81	26.50	58.64	41.91
Surprise	40.67	22.89	32.19	22.65	29.84	25.39	19.82	14.72
Fear	63.33	25.05	58.09	37.49	75.24	20.93	71.70	53.39
Disgust	21.33	19.87	10.09	12.82	26.67	20.55	16.19	17.85
Neutral	51.83	26.12	44.95	18.49	70.16	26.74	65.90	31.13
				Young Adults				
Happiness	57.85	19.17	51.83	27.15	46.72	21.61	39.11	19.52
Sadness	65.38	23.33	60.43	36.42	63.44	24.76	58.22	40.25
Anger	42.00	23.34	33.71	24.04	56.89	25.84	50.73	36.62
Surprise	39.08	27.26	30.38	20.30	37.87	26.57	28.99	20.68
Fear	66.31	25.03	61.50	40.08	73.44	19.66	69.65	53.85
Disgust	15.85	16.65	3.83	10.27	17.54	17.74	5.76	10.98
Neutral	52.62	26.84	45.85	18.84	66.07	24.65	61.22	26.33

***Note*. Children (5-10 years), Adolescents (11-19 years), Young Adults (20-39 years). ACC = Accuracy in Percent; SD_{ACC} = Standard deviation for accuracy; CCR = chance-corrected recognition in Percent; H_u = unbiased hit rates in Percent.**

Table 12: Overall accuracy (ACC), chance-corrected recognition rates (CCR), and unbiased hit rates (H_u) for age group and speaker sex (averaged across emotion categories and listener sex)

	Children	Adolescents	Young Adults	All Age Groups
		Female Speakers		
Accuracy	31.74	47.57	48.44	42.58
CCR	21.99	40.08	41.08	34.38
H_u	11.45	24.76	25.30	20.50
		Male Speakers		
Accuracy	34.88	52.81	51.71	46.47
CCR	25.58	46.07	44.81	38.82
H_u	14.24	31.10	29.75	25.03
		Female and Male Speakers Combined		
Accuracy	33.31	50.19	50.08	44.53
CCR	23.78	43.08	42.94	36.60
H_u	12.84	27.93	27.52	22.76

***Note*. Children (5-10 years), Adolescents (11-19 years), Young Adults (20-39 years). CCR = chance-corrected recognition; H_u = unbiased hit rates. All values in Percent.**

The CCRs of 42.94% presented for the adult listener group in this study are very similar to the 39% reported by Lassalle et al. (2019) in their UK sample, despite using socially-relevant word rather than sentence stimuli. The CCR values in both Lassalle et al. (2019) and this study appear fairly low in comparison to the 63% of CCR values reported in face research (O'Reilly et al., 2016) highlighting the difficulties recognising emotion accurately from speech.

Overall accuracy rates reported here are slightly lower than values reported elsewhere. For example, Morningstar, Ly, et al. (2018) reported 56.0% emotion recognition accuracy[4] for the youth listeners and 60.4% for adult listeners of adults speech samples. Similarly, Juslin and Laukka (2001) reported 56% overall recognition accuracy for a young adult listener group.

[4] Calculated from Table 2 in Morningstar et al. (2018), excluding Friendliness and Meanness

3.7 Supplementary material 3

Table 13: Pairwise comparison of emotion contrasts in predicted age trends

Emotion Contrasts	Estimate	Standard Error	Z ratio	P value
Happiness – Sadness	-0.032	0.006	-4.997	<.001
Happiness – Anger	-0.004	0.006	-0.578	.997
Happiness – Surprise	-0.013	0.007	-1.792	.554
Happiness – Fear	-0.041	0.007	-5.727	<.001
Happiness – Disgust	0.029	0.008	3.717	.004
Happiness – Neutral	-0.032	0.007	-4.737	<.001
Sadness – Anger	0.028	0.006	4.736	<.001
Sadness – Surprise	0.019	0.006	3.085	.033
Sadness – Fear	-0.009	0.007	-1.394	.805
Sadness – Disgust	0.061	0.007	8.716	<.001
Sadness – Neutral	-0.0002	0.007	-0.026	1.000
Anger – Surprise	-0.009	0.006	-1.477	.759
Anger – Fear	-0.037	0.007	-5.736	<.001
Anger – Disgust	0.033	0.007	4.662	<.001
Anger – Neutral	-0.028	0.007	-4.296	<.001
Surprise – Fear	-0.028	0.006	-4.386	<.001
Surprise – Disgust	0.042	0.007	5.813	<.001
Surprise – Neutral	-0.019	0.007	-2.728	.091
Fear – Disgust	0.070	0.007	9.883	<.001
Fear – Neutral	0.009	0.007	1.314	.846
Disgust – Neutral	-0.061	0.008	-7.925	<.001

***Note*. The estimate is the difference of predicted age trends (Table 7) between the compared emotion categories. Df = inf for asymptotic test.**

3.8 Supplementary material 4

Table 14: Contrast comparisons for model-based predicted probability between emotion categories

Emotion Contrasts	Estimate	Standard Error	Z ratio	P value
Happiness – Sadness	0.007	0.081	0.091	1.000
Happiness – Anger	-0.001	0.051	-0.015	1.000
Happiness – Surprise	0.238	0.063	3.766	.003
Happiness – Fear	-0.089	0.085	-1.053	.941
Happiness – Disgust	0.332	0.068	4.877	<.001
Happiness – Neutral	0.029	0.049	0.597	.997
Sadness – Anger	-0.008	0.081	-0.100	1.000
Sadness – Surprise	0.231	0.049	4.682	<.001
Sadness – Fear	-0.097	0.066	-1.470	.763
Sadness – Disgust	0.324	0.030	10.723	<.001
Sadness – Neutral	0.022	0.064	0.344	1.000
Anger – Surprise	0.239	0.058	4.094	<.001
Anger – Fear	-0.088	0.055	-1.620	.669
Anger – Disgust	0.332	0.063	5.318	<.001
Anger – Neutral	0.030	0.044	0.678	.994
Surprise – Fear	-0.327	0.050	-6.542	<.001
Surprise – Disgust	0.093	0.031	3.060	.036
Surprise – Neutral	-0.209	0.029	-7.244	<.001
Fear – Disgust	0.421	0.049	8.623	<.001
Fear – Neutral	0.118	0.052	2.265	.261
Disgust – Neutral	-0.302	0.043	-7.027	<.001

***Note*. All comparisons are tukey-corrected. The estimate is the difference of predicted probability to respond correctly between the compared emotion categories. Df = inf for asymptotic test.**

3.9 Supplementary material 5

The tables within Supplementary material 5 show contrast comparisons between emotion categories for children, adolescents, and young adults with female (Table 15) and male speakers (Table 16). These values were obtained from the analysis of the “listener age group” model.

Table 15: Differences in model-based predicted probability of correct response for emotion category contrasts from female speakers by listener age group

Emotion Contrast	Children (5-10 years)				Adolescents (11-19 years)				Young Adults (20-39 years)			
	Estimate	Standard Error	Z ratio	P value	Estimate	Standard Error	Z ratio	P value	Estimate	Standard Error	Z ratio	P value
HAP – SAD	0.106	0.119	0.889	.974	0.071	0.111	0.643	.995	-0.070	0.108	-0.647	.995
HAP – ANG	0.229	0.077	2.976	.046	0.273	0.075	3.655	.005	0.232	0.076	3.049	.037
HAP – SUR	0.324	0.092	3.524	.008	0.301	0.087	3.455	.010	0.306	0.090	3.407	.012
HAP – FEA	0.258	0.113	2.283	.252	0.017	0.112	0.157	1.000	-0.026	0.114	-0.226	1.000
HAP – DIS	0.394	0.098	4.029	.001	0.496	0.100	4.970	<.001	0.541	0.101	5.359	<.001
HAP – NEU	0.308	0.077	4.019	.001	0.158	0.065	2.432	.185	0.109	0.067	1.640	.656
SAD – ANG	0.123	0.105	1.177	.903	0.202	0.111	1.823	.532	0.301	0.104	2.886	.060
SAD – SUR	0.218	0.072	3.042	.038	0.230	0.073	3.162	.026	0.375	0.062	6.041	<.001
SAD – FEA	0.152	0.089	1.716	.605	-0.054	0.091	-0.588	.997	0.044	0.081	0.543	.998
SAD – DIS	0.288	0.047	6.116	<.001	0.425	0.041	10.360	<.001	0.611	0.037	16.577	<.001
SAD – NEU	0.202	0.085	2.387	.204	0.087	0.092	0.953	.964	0.179	0.083	2.168	.313
ANG – SUR	0.095	0.072	1.322	.842	0.028	0.088	0.315	1.000	0.074	0.089	0.835	.981
ANG – FEA	0.029	0.062	0.467	.999	-0.256	0.074	-3.474	.009	-0.257	0.074	-3.477	.009
ANG – DIS	0.165	0.076	2.154	.321	0.223	0.097	2.306	.241	0.310	0.092	3.371	.013
ANG – NEU	0.079	0.056	1.418	.792	-0.115	0.067	-1.717	.605	-0.122	0.067	-1.819	.535
SUR – FEA	-0.066	0.063	-1.053	.941	-0.284	0.077	-3.665	.005	-0.331	0.075	-4.436	<.001
SUR – DIS	0.070	0.043	1.613	.674	0.195	0.056	3.494	.009	0.235	0.044	5.306	<.001
SUR – NEU	-0.015	0.036	-0.425	1.000	-0.143	0.053	-2.694	.100	-0.196	0.053	-3.737	.004
FEA – DIS	0.136	0.062	2.202	.294	0.478	0.075	6.376	<.001	0.567	0.066	8.641	<.001
FEA – NEU	0.050	0.060	0.847	.980	0.141	0.076	1.855	.511	0.135	0.075	1.795	.551
DIS – NEU	-0.085	0.053	-1.602	.681	-0.338	0.074	-4.562	<.001	-0.432	0.067	-6.483	<.001

***Note*. All comparisons are tukey-corrected. HAP = Happiness, SAD = Sadness, ANG = Anger, SUR = Surprise, FEA = Fear, DIS = Disgust. Df = inf.**

Table 16: Differences in model-based predicted probability of correct response for emotion category contrasts from male speakers by listener age group

Emotion Contrast	Children (5-10 years)				Adolescents (11-19 years)				Young Adults (20-39 years)			
	Estimate	Standard Error	Z ratio	P value	Estimate	Standard Error	Z ratio	P value	Estimate	Standard Error	Z ratio	P value
HAP – SAD	0.080	0.092	0.868	.977	-0.011	0.112	-0.101	1.000	-0.116	0.113	-1.031	.947
HAP – ANG	-0.207	0.069	-3.001	.043	-0.312	0.075	-4.146	.001	-0.272	0.076	-3.577	.006
HAP – SUR	0.182	0.075	2.427	.187	0.204	0.090	2.268	.259	0.105	0.090	1.161	.909
HAP – FEA	-0.125	0.105	-1.190	.898	-0.318	0.109	-2.920	.054	-0.313	0.112	-2.802	.075
HAP – DIS	0.166	0.086	1.925	.464	0.130	0.111	1.177	.903	0.240	0.104	2.306	.241
HAP – NEU	-0.015	0.065	-0.233	1.000	-0.232	0.077	-3.003	.043	-0.185	0.077	-2.386	.204
SAD – ANG	-0.287	0.092	-3.123	.030	-0.301	0.093	-3.248	.020	-0.155	0.098	-1.578	.697
SAD – SUR	0.102	0.053	1.938	.455	0.215	0.070	3.091	.033	0.221	0.072	3.065	.035
SAD – FEA	-0.205	0.078	-2.647	.112	-0.307	0.082	-3.752	.003	-0.197	0.084	-2.338	.226
SAD – DIS	0.086	0.033	2.597	.127	0.142	0.045	3.166	.026	0.356	0.045	7.876	<.001
SAD – NEU	-0.095	0.069	-1.385	.810	-0.221	0.084	-2.625	.118	-0.068	0.088	-0.776	.987
ANG – SUR	0.389	0.071	5.475	<.001	0.516	0.062	8.276	<.001	0.376	0.072	5.250	<.001
ANG – FEA	0.082	0.073	1.119	.922	-0.006	0.058	-0.108	1.000	-0.042	0.065	-0.635	.996
ANG – DIS	0.373	0.084	4.457	<.001	0.442	0.090	4.893	<.001	0.512	0.083	6.133	<.001
ANG – NEU	0.192	0.065	2.951	.050	0.079	0.055	1.457	.770	0.087	0.062	1.410	.797
SUR – FEA	-0.307	0.063	-4.891	<.001	-0.522	0.057	-9.160	<.001	-0.418	0.063	-6.579	<.001
SUR – DIS	-0.016	0.045	-0.359	1.000	-0.074	0.069	-1.071	.937	0.136	0.054	2.491	.162
SUR – NEU	-0.197	0.038	-5.186	<.001	-0.436	0.045	-9.747	<.001	-0.289	0.051	-5.693	<.001
FEA – DIS	0.291	0.071	4.112	.001	0.449	0.081	5.556	<.001	0.553	0.069	8.021	<.001
FEA – NEU	0.110	0.071	1.557	.710	0.086	0.061	1.406	.799	0.129	0.068	1.904	.477
DIS – NEU	-0.181	0.060	-3.006	.042	-0.363	0.082	-4.437	<.001	-0.425	0.071	-5.971	<.001

***Note*. All comparisons are tukey-corrected. HAP = Happiness, SAD = Sadness, ANG = Anger, SUR = Surprise, FEA = Fear, DIS = Disgust. Df = inf.**

Chapter 4 The Glasgow vocal emotion and personality corpus

4.1 Introduction

Assessing information about a speaker quickly is essential for successful social communication. Receivers do not only rely on content being communicated, but are able to extract non-linguistic cues, such as age (Demenescu et al., 2014; Lima et al., 2014), sex (Schvartz & Chatterjee, 2012), identity (Lavan, Knight, et al., 2019), affect (Banse & Scherer, 1996; Juslin et al., 2018), and personality (Baus et al., 2019; Oleszkiewicz et al., 2017) when making speedy judgements even after brief exposure (Mahrholz et al., 2018; McAleer et al., 2014). The most crucial cues for effective social interaction are assessing trait and state characteristics, i.e. the speaker's personality and the current emotional state. Whether these impressions are accurate or not, they influence our immediate actions and behaviours towards others as to whether approaching or avoiding them (McAleer et al., 2014).

To study personality and emotion perception effectively, we need validated databases. There is a wealth of literature dedicated to vocal emotion processing which is also reflected in the number of databases and corpora dedicated to emotion. Partially influenced by the reproducibility crisis, many researchers are making their databases freely available whereas others have created corpora that can be requested from the authors for scientific use. Table 17 presents a selected overview of these rich corpora used in psychological research. As can be seen, most of the affect databases either include non-verbal vocalisations/ affect bursts (Belin et al., 2008; Lima et al., 2013; Sauter, Eisner, Ekman, & Scott, 2010; Sauter & Scott, 2007) or longer speech segments such as phrases or sentences (Burkhardt et al., 2005; Castro & Lima, 2010; Lassalle et al., 2019; Laukka et al., 2010; Laukka et al., 2013; Schirmer et al., 2019). Some multi-modal corpora containing separate vocal and facial stimuli are also included in Table 17, yet again they present either sentences or vowel sounds (Bänziger et al., 2009; Bänziger et al., 2012). However, an open-access database of socially-relevant words, which we frequently encounter in daily life, is missing. Creating a database of speech stimuli with high ecological validity would be beneficial for the field of vocal emotion research.

In contrast to the emotion databases, vocal personality has received less attention in the literature, which explains the limited number of open-access or upon-request databases and corpora dedicated to vocal personality traits (see Table 17). Currently, there are three with similar stimulus types. The Jena Speaker Set (JESS; Zäske et al., 2020) contains a multitude of different stimuli recordings from vowels to sentences, and semi-spontaneous speech. The Geneva Faces and Voices database (GEFAV; Ferdenzi et al., 2015) presents vowels and sentences, and Mahrholz et al. (2018) include ambiguous and socially-relevant word and sentence stimuli. All three databases are validated on various different social traits. The stimuli by Mahrholz et al. (2018) and in the GEFAV (Ferdenzi et al., 2015) are both validated on trustworthiness, dominance, and on attractiveness. The GEFAV adds further social traits such as masculinity/ femininity or health. The JESS (Zäske et al., 2020) also includes attractiveness ratings, but does not include the two key personality traits of trustworthiness and dominance that were identified in the social voice space model (McAleer et al., 2014). Yet, the stimuli were validated on likeability which has been shown to be closely related to trustworthiness (McAleer et al., 2014).

In relation to speaker recruitment, Ferdenzi et al. (2015) and Mahrholz et al. (2018) recruited only younger adult encoders, though sample sizes differ substantially with Mahrholz et al. (2018) including 60 speakers, and the GEFAV (Ferdenzi et al., 2015) holding samples from 111. However, the JESS (Zäske et al., 2020) expands on speakers age to include younger and older adult speakers with approximately 60 speakers per age group. Finally, a distinct feature between the three databases is stimulus language. The GEFAV (Ferdenzi et al., 2015) is in French, the JESS (Zäske et al., 2020) in German, and the stimuli set by Mahrholz et al. (2018) is in English. Selecting stimuli similar or different to the listeners' native language may influence perceptions (Giles & Billings, 2004). This shows that much variability exists between the three corpora and all have different strengths and shortcomings in relation to one another. The majority of representations in the three databases, bar semi-spontaneous speech, are emotionally neutral which may result in low ecologic validity.

Table 17: Selected overview of recent open-access datasets with enacted categorical emotion dimensions and/or personality ratings; typically used in psychological research

Database	Stimulus type	Speakers/ items	Rating scales	Available
Corpora of affective non-verbal vocalisations				
Sauter (e.g. Sauter, Eisner, Ekman, & Scott, 2010; Sauter & Scott, 2007)	Non-verbal vocalisations, laughter (various sets)	Various (from France, Japan, Namibia, the Netherlands, United Kingdom and the United States)	Anger, fear, disgust, sadness, surprise, amusement, triumph, relief, contempt, embarrassment, guilt, shame, awe, compassion, contentment, desire, enthusiasm, gratitude, interest, love, pride, sensory pleasure, laughter	Upon request https://aice.uva.nl/research-tools/research-tools.html
Montreal Affective Voices (MAV; Belin et al., 2008)	Affect bursts (no interjections; limited to vowel sound 'ah')	10 Francophone actors (5F)	Angry, disgusted, fearful, happy, painful, pleased, sad, surprised, neutral, valence, arousal, intensity	Open-access https://neuralbasesofcommunication.eu/download/
Lima et al. (2013)	Non-verbal vocalisations (no interjections not limited to specific vowel sounds)	4 European Portuguese native speakers without formal acting training (2F, 27 and 33 years; 2M, 28 and 34 years)	Achievement/triumph, amusement, sensual pleasure, relief, anger, disgust, fear, sadness, valence, arousal, authenticity	Open-access https://link.springer.com/article/10.3758/s13428-013-0324-3

Corpora of affective phrases, sentences, and longer speech segments				
Berlin Database of Emotional Speech (Emo-DB; Burkhardt et al., 2005)	German sentences with neutral meaning	10 actors (5F, age range 21-35 years)	Neutral, anger, fear, joy, sadness, disgust, boredom	Open-access http://emodb.bilderbar.info/index-1280.html
Castro and Lima (2010)	Portuguese sentences and pseudo-sentences	2 women (mean age = 18 years) with musical training	Neutral, anger, disgust, fear, happiness, sadness, surprise, intensity[5]	Open-access https://link.springer.com/article/10.3758%2FBRM.42.1.74
Vocal Expressions of Nineteen Emotions across Cultures (VENEC; Laukka et al., 2010; Laukka et al., 2013)	Verbal materials: short neutral phrases, and longer paragraph of neutral text; non-linguistic vocalisations (subset of actors only)	100 professional actors (50F; age range 5-30 years) from 5 English-speaking cultures (USA, India, Kenya, Singapore, Australia)	Affection, amusement, anger, contempt, disgust, distress, fear, guilt, happiness, interest, lust, negative surprise, neutral, positive surprise, pride, relief, sadness, serenity, shame	Upon request (petri.laukka@psychology.su.se)
Schirmer et al. (2019)	English sentences with neutral meaning in: 6 emotions (content, happy, proud, afraid, angry, sad); 4 conversational expressions (confident, stating, doubtful, questioning); enacted as trustworthy, untrustworthy, and neutral versions	20 Singaporean native English amateur-actors: 10 young (5F, mean age = 22.2 years; 5M, 23.8 years), and 10 older (5F, 69.2 years; 5M, 63.0 years)	Valence and arousal (preliminary study); perceived trustworthiness	Open-access https://osf.io/j3hfg/

[5] Intensity is described as “how representative the stimulus was of the chosen category”, which fits more with the terminology of “recognisability” in this validation study.

The EU-Emotion Voice Database (Lassalle et al., 2019)	English sentences (some congruous, some neutral)	18 (9F) per language (English, Swedish, Hebrew)	Afraid, angry, ashamed, bored, disappointed, disgusted, excited, frustrated, happy, hurt, interested, jealous, joking, kind, proud, sad, sneaky, surprised, unfriendly, worried, neutral, valence, arousal, intensity	Upon request https://www.autismresearchcentre.com/tests/the-eu-emotion-stimulus-set/
Multi-modal affect corpora (including auditory-only dimensions)				
Multimodal Emotion Recognition Test (MERT; Bänziger et al., 2009)[6]	Auditory-only modality: two pseudo-sentences	12 German-speaking professional actors (6F); only 10 included in final set	Irritation, anger, anxiety, fear, happiness, elated joy, disgust, contempt, sadness, despair	Upon request https://www.unige.ch/cisa/emotional-competence/home/research-tools/mert/
GEneva Multimodal Emotion Portrayals (GEMEP; Bänziger et al., 2012)[7]	Auditory-only modality: two pseudo-sentences (statement, question), vowel sounds 'aaa')	10 French-speaking theatre actors (5F; mean age = 37.1 years, age range 25-57 years)	Admiration, amusement, tenderness, hot anger (rage), disgust, despair, pride, anxiety, interest, irritation (cold anger), elated joy, contempt, (panic) fear, pleasure relief, surprise, sadness, neutral	Upon request https://www.unige.ch/cisa/gemep

[6] Recordings taken from the GVEESS corpus (Banse & Scherer, 1996).

[7] The ERAM (Laukka et al., 2021) is considered the short version of the MERT and consists of a series of 72 brief audio-video (24 audio-only) taken from the GEMEP corpus (Bänziger et al., 2012). It is therefore not listed separately, however, stimuli can also be requested via https://www.unige.ch/cisa/emotional-competence/home/research-tools/.

Personality corpora				
Geneva Faces and Voices database (GEFAV; Ferdenzi et al., 2015)	Auditory stimuli: Three consecutive vowels [/i/, /a/, /o/] and sentences in French with neutral meaning	111 French speakers (61F; age range 18-35 years)	Voice ratings only: Attractiveness, trustworthiness, dominance (study 2), masculinity/ femininity, beauty (study 1), and health (study 2)	Upon request https://www.unige.ch/cisa/gefav
Mahrholz et al. (2018)	English words and sentences with socially-relevant and ambiguous meaning	60 students from Scotland (30F, mean age = 20.2 years, age range 17-27 years; 30M, 23.2 years, age range 17–30 years)	Perceived trustworthiness, dominance, attractiveness	Open-access https://osf.io/s3cxy/
Jena Speaker Set (JESS; Zäske et al., 2020)	German sentences, syllables, read text, semi-spontaneous speech, and vowels	120 German speakers: 61 young (30F; 18–25 years), and 59 old (29F; 60–81 years)	Perceived attractiveness, likeability, distinctiveness, regional accent, and age	Open-access https://osf.io/u6amw/

As can be seen from Table 17, there are emotive databases that include many emotion categories from a variety of speakers, but lack personality ratings. Likewise, the majority of stimuli in the personality corpora are validated on emotionally-neutral read-out scenarios that do not include affect representations. However, speech in the real-world is hardly ever truly unemotive. There is a gap in the literature for an affect corpus that also includes personality assessments from the same speakers to increase ecological validity. Here, we are aiming to produce an extensive validated vocal emotion corpus that integrates both personality and emotion concepts, is openly accessible (with a Creative Commons Attribution 4.0 International License; CC BY 4.0), free of charge, and can be used in a variety of settings, such as for research or teaching.

The starting point for this database is the frequently-used, semantically-neutral, socially-relevant word “hello” allowing us to expand on the already existing affect speech corpora of words and sentences. In the following, we present 312 stimuli from 24 native Scottish English speakers (12 female) who encoded the socially-relevant word “hello” in the basic 6 emotion categories of happiness, sadness, anger, surprise, fear, and disgust (as suggested by Ekman & Friesen, 1971; Sauter, Eisner, Ekman, & Scott, 2010), both in a high- and low-intensity version, plus one neutral representation. The stimuli are validated on trait ratings of trustworthiness, dominance, and attractiveness (similar to Ferdenzi et al., 2015; Mahrholz et al., 2018), perceived emotion category, recognisability and authenticity ratings (as shown in Banse & Scherer, 1996; Morningstar et al., 2017; Morningstar, Ly, et al., 2018), valence, arousal, and perceived intensity ratings (similar to Belin et al., 2008). We will also pay close attention to listener sex differences to justify the values reported in the database.

Section 4.2 focuses on the creation of the vocal stimuli that were subsequently used in Studies 1 to 4. Validation of the stimuli on trustworthiness, dominance, and attractiveness is discussed in Study 1 (Section 4.4). Perceived emotion, recognisability, and authenticity are investigated in Study 2 (Section 4.5). Subsequently, Study 3 (4.6) explores valence and arousal perceptions, whereas perceived intensity is validated in Study 4 (4.7). Whilst acoustic measures will be provided as part of the database and are getting briefly outlined in section 4.2.6

below, they were not further analysed within the scope of this thesis, and therefore, will not be specifically addressed in a separate study. Free response data were also collected. Since data analysis is still ongoing, it is excluded from the thesis, however, information will be added as Study 5 in a future publication.

The Glasgow vocal emotion and personality corpus is available on OSF https://osf.io/6da4r/.

4.2 Vocal stimuli creation

4.2.1 Speakers

Twenty-four native English speakers were recruited for stimuli recording via the University of Glasgow School of Psychology Subject Pool (see Table 18 for demographic information). Advertisement criteria included participants who have grown up in Scotland, were under the age of 40 on the day of recording, had no speech impediments, and either had experience in acting (e.g. voice acting, radio advertising, etc.) or would consider themselves comfortable producing recognisable voice stimuli in a recording booth. Recording sessions lasted around 1 to 1h30min, and speakers received £15 for their contribution.

Table 18: Demographic information of speakers

Speaker Sex	N	Mean Age	SD Age	Min Age	Max Age
Female	12	21.6	1.88	19	24
Male	12	22.3	4.60	18	34

A Welch two-sample t-test revealed no significant differences of speaker age between the female and male speakers, $t(14.58) = 0.523$, $p = 0.609$.

4.2.2 Questionnaires and recording materials

A demographics questionnaire contained questions about the speaker that may be related to physical properties of the voice: sex, sexuality, nationality, accent, ethnicity, date of birth, country of birth, city/province of birth, height,

weight, hearing difficulties, first language, other languages, and whether they are a smoker.

The following tasks were recorded by each speaker on the recording day:

1. Free speech: An impromptu scenario starting off with a “radio show” in which the researcher tried to elicit the word “hello” from the “guest speaker” in the most natural way. Speakers were also asked to describe directions between two landmarks either around the Glasgow Westend, or in case they were unfamiliar to the area, in the Glasgow City Centre. Furthermore, they also described their favourite dish and how to prepare it.

2. Creative speech: An impromptu scenario in which speakers were asked to create a story using the words: “hello”, “colours”, “left”, and “right”.

3. Reading task 1: Speakers were instructed to read the following passages a) in a natural way without emotional intonation, and b) as a story book teller (Rainbow Passage and Telephone scenarios) or as if saying something to a friend/ family member (Harvard Sentences).

    I. An abridged version of the Rainbow Passage (Fairbanks, 1960; see Supplementary material 1)

    II. Two telephone scenarios that have been previously used by our lab (see Supplementary material 1)

    III. Two sets of the Harvard Sentences ("IEEE Recommended Practice for Speech Quality Measurements," 1969; see Supplementary material 1)

4. Reading task 2: A selected list of short words to be read in a neutral/ unemotional way (see Supplementary material 1)
5. Affect reading task: Speakers were instructed to read the following stimuli in the emotion categories happy, sad, angry, surprised (positive), fearful, and disgusted in a) a subtle/ non-theatrical (i.e. low-intensity

condition), and b) theatrical way (i.e. high-intensity condition). Speakers also encoded a non-emotional/ neutral expression.

I. Vowels (a [/eɪ/], e [/iː/], i [/ʌɪ/], o [/əʊ/], u [/juː/])

II. Numbers 1 to 10

III. Days of the week

IV. Words: “hello”, “colours”, “left”, and “right”

V. Pseudo-Sentences (Ethofer et al., 2009; Pell et al., 2009; Rigoulot et al., 2013; see Supplementary material 1)

6. Acting task: non-verbal expressions of emotion categories happiness, sadness, anger, surprise (positive), fear, disgust, and neutral (Belin et al., 2008)

Audacity Version 2.3.0 (Audacity Team, 2021) was used to record (.wav format, 16-bit mono, 44100 Hz) the stimuli.

4.2.3 Recording procedure

The recording materials (apart from the impromptu parts 1 and 2) with detailed instructions were sent to the voice actors a few days before their recording session so they could familiarise themselves with the content. This was done to allow speakers to produce the stimuli in a more natural way. Speakers were invited to record their reading and acting attempts in a custom-made sound-attenuated chamber within the School of Psychology at the University of Glasgow. On the day of recording, speakers would provide written consent, before filling in the demographics questionnaire and start the recording session.

Speakers recorded all materials standing, and were allowed to take breaks and drink water between segments. The order of recordings was similar for each participant: Speakers always started with the free speech and creative speech scenarios, followed by reading task 1, reading task 2, and the affect reading task. The acting task was always last.

During the affect reading task, each speaker was required to encode the affect speech scenarios in a low- and high- emotional intensity representation per emotion category (i.e. happiness, sadness, anger, positive surprise, fear, disgust). Speakers were allowed to choose which order they wanted to record materials in (e.g. all low-intensity emotion first in a given recording task; alternate between low- and high-intensity representations of the same emotion first; all vowels/numbers first, then words, then sentences; etc.). A neutral representation was only recorded once. When describing the affect scenarios to the speakers, we used the labels “subtle” and “theatrical” instead of low and high intensity respectively. This was done to avoid speakers producing the same emotion in a louder way rather than with higher emotional intensity. No emotive scenarios or vignettes were provided to allow the speakers to express each emotion category as they saw fit which is in line with recording procedures elsewhere (Schirmer et al., 2019). There was also no feedback provided as to whether the emotions produced were identifiable or not.

For the elicitation of non-verbal vocalisations in the acting task, speakers were allowed to record as many attempts as they felt necessary to produce the requested emotion category. Similar to the affect reading task, no instructions or feedback were provided to guide speakers as the aim was to elicit their interpretation of a given emotion category.

4.2.4 Ethics

All recording procedures were approved by the University of Glasgow College of Science and Engineering Ethics Committee (No. 300150058, 300170290) and are in accordance with the ethical standards of the 1964 Declaration of Helsinki. Recording participants (i.e. speakers) were consenting to providing their speech samples to an open-source database that would be freely available for researchers to use. They were allowed to skip any questions in the demographics questionnaire they would not want to provide an answer to. They were informed they had a right to withdraw at any stage of the recording process and would be paid pro rata for their commitment up until that point. Since the recording procedures also included an impromptu free-speech scenario in which participants were asked to greet a fictional audience, they were told that any

identifying personal information would be deleted from the recording before being shared with other researchers.

4.2.5 Pre-processing of vocal clips

As a starting point for a validated database, we focused on recordings of the socially-relevant word “hello” obtained during the affect recording task. In total, 312 stimuli [12 speakers x 2 speaker sex x (6 emotion categories x 2 intensity levels + 1 neutral representation)] were extracted from the recordings using Audacity Version 2.3.0 (Audacity Team, 2021). Audacity was also used to remove recording artifacts. Subsequently, MATLAB (Version 9.1 (R2016b); MATLAB, 2016) was used to normalise stimuli for sound intensity to account for potential differences in loudness between speakers (e.g. some speakers standing further away from the microphone than others). We opted to normalise the stimuli within each emotion category to maintain acoustic features specific to each of those emotion categories (Chen et al., 2012; Kamiloğlu et al., 2020; Schirmer et al., 2007).

4.2.6 Acoustic measures

Acoustic information was obtained from the 312 normalised stimuli. In line with previous work from our lab and in collaboration with others (e.g. Baus et al., 2019; McAleer et al., 2014), the following measures were obtained via Praat (Version 6.1.55; Boersma & Weenink, 2021):

1) Duration in ms

2) Mean fundamental frequency (f0; range: min 75 Hz; max: 600 Hz) relating to pitch.

3) Intonation calculated as the difference between maximum and minimum mean fundamental frequency ($f0_{max}$ – $f0_{min}$).

4) Glide calculated as the difference between mean fundamental frequency at the start and at the end of the stimulus ($f0_{end}$ – $f0_{start}$).

5) Formant dispersion as a measure of vocal tract size (ratio between formants means F1 to F5). Burg linear predictive coding algorithm was used; maximum formant frequency was set to 5500 Hz; window length was 0.025 sec.

6) Harmonic-to-noise ratio (HNR) indicating roughness in the voice, on the basis of forward cross-correlation analysis (mean value; time step = 0.01 sec; minimum pitch 75.0 Hz, silence threshold = 0.1; periods per window = 1.0).

7) Jitter as a measure of cycle-to-cycle variations in frequency (Patel et al., 2011; Teixeira et al., 2013), via Relative Average Perturbation (RAP). It is calculated as the average absolute difference between an interval and the average of that interval and its two neighbours (shortest period = 0.0001 sec; longest period = 0.02 sec; max. period factor = 1.3)

8) Shimmer as a measure of cycle-to-cycle variations in amplitude (Patel et al., 2011; Teixeira et al., 2013). This is the three-point Amplitude Perturbation Quotient (APQ3) calculated as the average absolute difference between a periods amplitude and the average of amplitudes of its neighbours, divided by the average amplitude (shortest period = 0.0001 sec; longest period = 0.02 sec; maximum period factor = 1.3; maximum amplitude factor = 1.6);

9) Alpha ratio as a measure of spectral balance computed from the long-term average spectrum (LTAS). It is calculated as the ratio of mean energy within low (0–1000 Hz) and high frequencies (1000–5000 Hz) as suggested by Kitzing (1986).

10) Loudness measured in db.

4.3 General methods for studies 1-4

4.3.1 Ethics

All validation procedures were approved by the University of Glasgow College of Science and Engineering Ethics Committee (No. 300150058, 300170290) and are

in accordance with the ethical standards of the 1964 Declaration of Helsinki. Rating participants (i.e. Listeners) were informed their data would contribute to the validation of an open-access vocal emotion and personality database. They were informed that they could withdraw from the study at any time by simply closing the browser, and that none of the data provided up until the point of withdrawal would be contributing to the analysis.

4.3.2 Analysis – packages and environment

The following applies to studies 1 to 4: All data wrangling, visualisations, and analysis was completed in R (Version 4.0.4; R Core Team, 2020) and RStudio (Version 1.1.463; RStudio Team, 2016). Packages used: tidyverse (Version 1.3.1; Wickham et al., 2019), rstatix (Version 0.7.0; Kassambara, 2021), car (Version 3.0-11; Fox et al., 2021), psych (Version 2.1.6; Revelle, 2021), lubridate (Version 1.7.10; Grolemund & Wickham, 2011), hms (Version 1.1.0; Müller, 2021), irrNA (Version 0.2.2; Brückl & Heuer, 2021), Hmisc (Version 4.5.0; Harrell Jr, 2021), lme4 (Version 1.1-27.1; Bates et al., 2015), ordinal (Version 2019.12-10; Christensen, 2015), optimx (Version 2021-6.12; Nash & Varadhan, 2011), and emmeans (Version 1.6.3; Lenth, 2021).

4.4 Study 1: Vocal stimuli validation on personality traits of trustworthiness, dominance, and attractiveness

4.4.1 Methods

4.4.1.1 Power analysis

The sample size for detecting a main effect of listener sex was calculated using the power analysis tool PANGEA (https://jakewestfall.shinyapps.io/pangea/) designed by Jake Westfall (2016). Previous research from our lab has produced very small non-significant effect sizes between male and female listeners for perceived trustworthiness, dominance, and attractiveness. Hence, assuming an effect size of 0.45 as recommended by Westfall (2016), we would require a minimum of 18 participants per rating scale (trustworthiness, dominance, and attractiveness) and listener sex to achieve a power of .9.The recruitment numbers in this study were achieved and the study is therefore sufficiently powered.

4.4.1.2 Listeners

In total, 256 participants were recruited via the University of Glasgow School of Psychology Subject Pool to partake in an experiment on vocal trait ratings of either trustworthiness, dominance, or attractiveness. Recruitment criteria stipulated that participants had to be between 16 and 39 years of age, did not have any hearing impairments or hearing aids, were not currently taking any medication related to a mental health condition (e.g. anti-depressants, anti-anxiety), or had their voice recorded for the database.

Participants' contributions were excluded if they violated the initial recruitment criteria, had not rated all 312 stimuli, took part in more than one trait experiment, or when they rated the same stimulus multiple times (e.g. technical error or timed out and restarted experiment). In the latter two instances, the initial judgements were kept, and subsequent ratings were deleted. Initially, it had been intended that participants should not have had any prior exposure to the particular voice stimuli, however difficulties in recruitment during the COVID-19 pandemic led to slight adjustments. Decisions were made to include participants' trait data if the timeframe between taking part in the emotion and trait experiments exceeded 30 days (n = 3; average time = 123.01 ± 72.34 days). In total, 98 participants were excluded (39 for trustworthiness, 30 for dominance, 29 for attractiveness) resulting in a final sample size of 158 listeners (mean age = 22.5 ± 4.41 years, range: 17-39). Demographic information per listener sex and trait rated can be found below in Table 19.

Table 19: Demographic information of participants in the trait rating experiments

Listener Sex	N	Mean Age	SD Age	Min Age	Max Age
Trustworthiness					
Female Listeners	26	21.9	3.57	18	35
Male Listeners	26	22.2	4.75	17	39
Non-binary Listeners	1	17.0	NA	17	17
NA	1	19.0	NA	19	19
Dominance					
Female Listeners	26	24.0	4.51	18	35
Male Listeners	26	21.6	3.67	17	31
Attractiveness					
Female Listeners	26	22.5	3.91	18	31
Male Listeners	26	23.3	5.62	18	36

The experiments took approximately 45 min to complete and participants were reimbursed by either £5 in cash (before the pandemic), a £5 Amazon voucher (during the pandemic) or 3 participation credits for students at the University of Glasgow as part of their year 1 undergraduate course requirements.

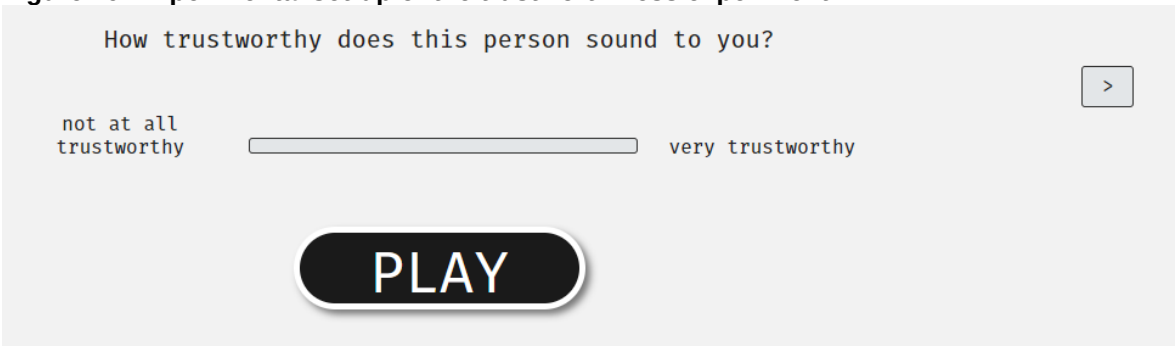
4.4.1.3 Materials, experimental set-up, and rating procedures

Experiments were created and hosted on the platform Experimentum (DeBruine et al., 2020). After having received general information about the purpose of the study, participants were informed that participation would be voluntary, their data would be stored and treated anonymously, and that they were allowed to withdraw from the study at any time by closing the browser. Participants provided consent via a yes/no option before answering a brief demographics questionnaire and partaking in either the trustworthiness, dominance or attractiveness experiment.

Experiments were designed to include 4 practice trials to familiarise participants with the experimental set-up and allow for adjustments of volume if necessary. During the experimental stage, participants would rate 312 vocal emotion representations across two consecutive blocks of 156 (one block per speaker sex). The order of blocks were counterbalanced and the stimuli within each

block randomised. During experimental trials, participants would see the question “How [trait] does this person sound to you?” above a visual analogue scale (VAS) slider ranging from “not at all [trait]” on the left to “very [trait]” on the right. They had to press a “play” button to hear a vocal emotion representation, make their selection on the slider, and confirm by pressing an arrow to get to the next trial (see Figure 19). Slider ratings could only be provided after the vocal stimulus was played in full.

Figure 19: Experimental set-up of the trustworthiness experiment



4.4.2 Results

4.4.2.1 Initial data preparation and reliability

Each of the 312 vocal affective stimuli was rated 52 times each on each trait (split evenly between female and male listeners). There were additional trustworthiness ratings from one non-binary participant and one listener who did not reveal their sex. The data of these two participants were included in the calculations determining overall mean trustworthiness scores, but excluded for the analysis of listener sex.

Slider values were recorded on a scale from 0 to 100. Responses were z-scored for each participant to avoid potential differences of scale use between participants. Subsequently, average scores for trustworthiness, dominance, and attractiveness were calculated from the z-scored ratings for each stimulus. Intraclass correlation coefficients (ICCs) as indices of interrater reliability and their 95% confidence intervals were computed on the z-scores of participants’ trustworthiness, dominance, and attractiveness ratings. We chose a two-way mixed-effect model, type consistency, that is based on the mean ratings of k raters since we have a fully crossed model, and our interest lies within the

raters of this study rather than generalising findings to raters with similar characteristics (Hallgren, 2012; Koo & Li, 2016; Shrout & Fleiss, 1979). Interrater reliability can be considered excellent for all three trait dimensions (Koo & Li, 2016; see Table 20 for details) which is comparable to previous studies (e.g. Baus et al., 2019; Mahrholz et al., 2018; McAleer et al., 2014; Schirmer et al., 2019; Zäske et al., 2020).

Table 20: Intraclass correlation coefficients with 95% confidence intervals for each of the perceived trait dimensions trustworthiness, dominance, and attractiveness

Perceived Trait Dimension	ICC	CI Lower Boundary	CI Upper Boundary
Trustworthiness	.941	.931	.950
Dominance	.954	.946	.961
Attractiveness	.961	.955	.967

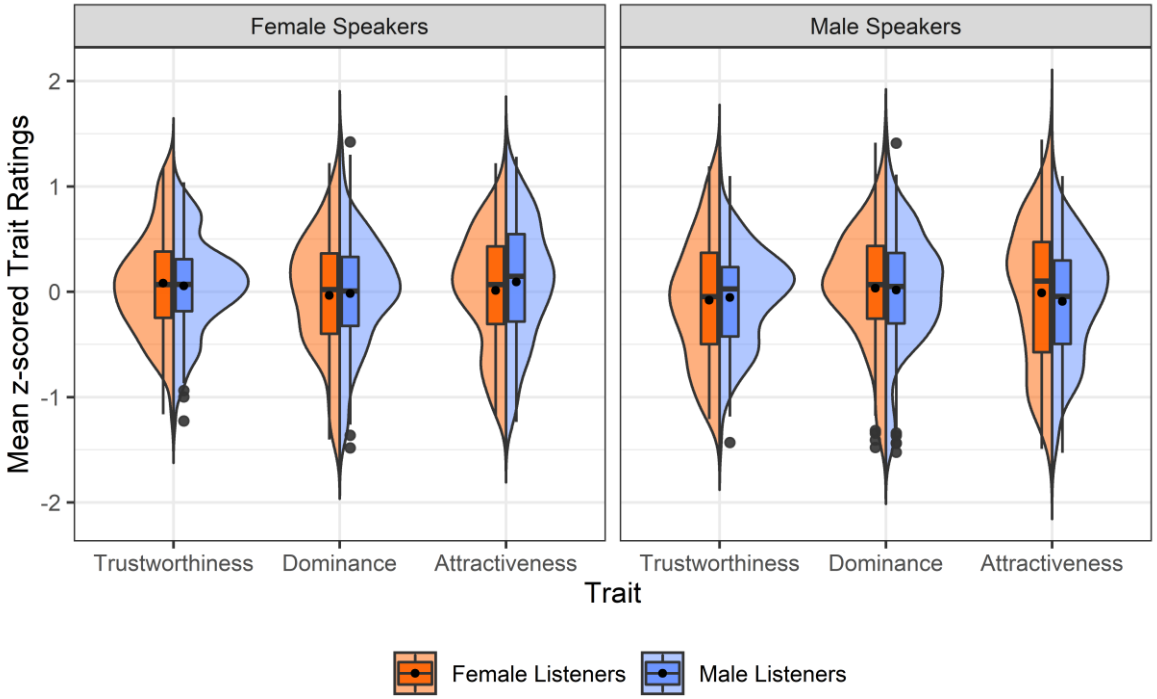
4.4.2.2 Listener sex differences

We computed a linear mixed-effects model with by-item and by-participant random slopes and intercepts (see model (1) below) to investigate differences between female and male listeners' ratings for each trait. The dependent variable was the raw trait ratings of either trustworthiness, dominance, or attractiveness. Speaker sex and the interaction of listener sex and speaker sex were added as control variables. Listener sex and speaker sex were deviation-coded and the optimizer optimx (Nash & Varadhan, 2011) was used to avoid conversion issues.

(1) Trait Ratings ~ Listener Sex * Speaker Sex + (1 + Speaker Sex | Listener ID) + (1+ Listener Sex | Speaker ID)

No significant differences were detected for listener sex, speaker sex, or interaction between the two, neither within trustworthiness, dominance, nor attractiveness (all $p > .05$; see Figure 20). Therefore, trait values for each of the 312 stimuli will be reported as a single score within the database, irrespective of listener sex.

Figure 20: Split violin-boxplots of mean z-scored trait ratings, separately by speaker sex and listener sex



Note. **Female listeners depicted on the left (orange) and male listeners on the right (blue) of each violin-boxplot. Within the Figure, z-scores were used to take individual scale use into account. For the mixed-effects model, raw rating scores were used since the random effects structure does account for differences of scale use behaviour between participants.**

4.4.3 Discussion

Female and male listeners did not significantly differ in their perceptions of the trait dimensions investigated within this study. This is in line with previous literature from our lab or other trait databases that did not observe significant main effects of listener sex either (Baus et al., 2019; Mahrholz et al., 2018; McAleer et al., 2014; Zäske et al., 2020). Contrarily, some studies have shown listener sex differences for attractiveness which has been argued to stem from an unwillingness of male listeners to rate male speakers on an attractiveness dimension (Babel et al., 2014). Yet, the authors also reported that overall, male and female listeners agreed on who sounded more/less attractive despite the lower ratings provided by the male listeners. Whilst not finding a main effect of listener sex for attractiveness, Zäske et al. (2020) reported a three-way interaction of listener sex with speaker sex and speaker age. It may be that listener sex accounts for small variations in relation to other variables such as age of the speaker. Yet, given that we did not find a main effect of listener sex or interaction with speaker sex, it is impossible for us to speculate. Validation values in the corpus are therefore reported irrespective of listener sex.

4.5 Study 2: Vocal stimuli validation on perceived emotion category, recognisability, and authenticity

4.5.1 Methods

4.5.1.1 Power analysis

PANGEA (Westfall, 2016) was used for calculating how many ratings each vocal stimulus would need to receive for a main effect of listener sex to be detected. A recent large-scale study (Lausen & Schacht, 2018) showed effect sizes for words to vary between 0.14 and 0.45 (phi coefficient r_φ recalculated from reported main effects) depending on the type of word used [0.45 for pseudo words, 0.37 for semantically positive nouns, 0.24 for semantically negative nouns, 0.25 for semantically neutral nouns, and 0.14 for affect bursts (ns)]. Since we are unsure which word type exclamations such as “hello” would be semantically closer to, we assumed a default effect size of 0.45 recommended by Westfall (2016). Twenty-eight ratings per voice stimulus per listener sex would be needed to achieve a power of .9. Given participants rated a subset of 39 male and 39 female vocal stimuli, the complete experiment required a minimum of 224 participants (112 female). The final sample size for the emotion rating experiment was therefore reasonably sensitive to the effects of interest.

4.5.1.2 Listeners

For the emotion validation experiment, 331 participants were recruited via the University of Glasgow School of Psychology Subject Pool. Participants’ contribution was removed when they did not provide all 39 ratings in each of the 6 rating blocks (i.e. 234 ratings in total), when they were 40 years of age or older, when they took part repeatedly (in that case, the data from the first rating session was kept), or when they selected one emotion category 50% of the time or more. This resulted in a final sample size of 256 participants with 139 self-identifying as female (mean age = 23.4 ± 4.27 years, range: 18-38), 114 as male (mean age = 22.4 ± 4.66 years, range: 17-36), and 2 as non-binary (mean age = 22.0 ± 1.41 years, range: 21-23). One person did not provide any demographic information on sex (age = 24 years).

Participants were allowed to take part if they were between 16 and 39 years of age, did not have any hearing impairments or hearing aids, were not currently taking any medication related to a mental health condition (e.g. anti-depressants, anti-anxiety), or had taken part in the voice recording procedure. Given the recruitment difficulties experienced during the pandemic, the decision was made to include participants if they had completed one of the trait experiments beforehand (n = 33). Contrary to trait impressions in which we aimed for zero acquaintance judgements, the initial impression formed about an emotional expression was secondary to the researchers.

Completion time was ca. 20 min, and participants were compensated by either £3 in cash (pre-pandemic), a £3 Amazon voucher (during the pandemic), or the equivalent in participation credits as course requirements for their undergraduate Psychology degree.

4.5.1.3 Materials, experimental set-up, and rating procedures

The experiment was created and hosted on the platform Experimentum (DeBruine et al., 2020). Participants were informed about their ethical rights, and the same consent-obtaining procedures were applied as outlined in the trait experiment. For this experiment, the 312 stimuli (156 per speaker sex) were divided into four subsets of 39 vocal stimuli per speaker sex. Two counterbalanced versions for each speaker sex were created to ensure that ratings did not depend on the particular combination of vocal clips in a subset. The following rules were applied when creating the subsets of 39 stimuli for each counterbalanced version:

1. Each subset contains 6 emotive stimuli per emotion category (i.e. happiness, sadness, anger, surprise, disgust, fear) and 3 neutral stimuli.

2. Each subset contains 3 subtle and 3 theatrical representations for each of the 6 emotive stimuli (i.e. in total, there are 18 subtle, 18 theatrical, and 3 neutral stimuli per subset).

3. Each subset contains 3 emotive representations of each speaker; the neutral representations from the speakers are distributed evenly across

the four subsets. Each speaker is therefore represented 3 or 4 times in each subset, depending on whether their neutral stimulus was added to that particular subset.

4. Each speaker in each subset is only represented once per intensity level per emotion type (i.e. if subset 1 already hosts "voice 1 anger low", "voice 1 anger high" is not selected within the same subset).

5. Each speaker is represented with a minimum of one low- and one high-intensity representation per subset (i.e. either 2 low + 1 high, or 2 high + 1 low)

At the beginning, participants were informed about the study's purpose, provided consent, and filled in a brief demographics questionnaire. They would complete three blocks of the same 39 representations per speaker sex: emotion recognition, recognisability, and authenticity. Speaker sex was counterbalanced and listeners would either rate the three blocks encoded by female or by male speakers first.

"Emotion recognition" was always the first block to be presented within each speaker sex as it required the listener's first impression judgment of emotion, whereas the order of "recognisability" and "authenticity" was counterbalanced. For "emotion recognition", participants could choose between 8 alternative forced choice (8-AFC) options (i.e. happiness, sadness, anger, surprise, fear, disgust, neutral, and other) to label what emotion they thought was represented in the recording. During "recognisability" and "authenticity", participants were provided with the label of the speaker's intended emotion category and asked to rate how recognisable and how authentic (i.e. genuine or real) the presented emotion was respectively. Recognisability and authenticity were rated on a 4-point Likert scale, with 1 = "not at all [*recognisable* or *authentic*]" to 4 = "very [*recognisable* or *authentic*]" similar to Morningstar et al. (2017; 2018). Listeners pressed a "play" button first to hear the stimulus. Selecting an emotion category confirmed the choice and progressed participant to the next rating screen. Response categories could only be selected after the vocal stimulus was played in full.

4.5.2 Results

4.5.2.1 Initial data preparation and reliability

Perceived emotion category. It was difficult to gather exactly the same number of listeners per subset on the experimental platform used. Participant numbers and therefore number of responses for each vocal stimulus varied per subset and speaker sex. This resulted in each of the 312 vocal stimuli receiving between 57 and 73 ratings with a minimum contribution of 28 female and 28 male listeners (see https://osf.io/6da4r/ for detailed information).

To determine mean accuracy for the perceived emotion ratings, "correct response" was scored as 1 when perceived emotion category selected by the listener matched the speaker's intended response category. All other response options and the "other" category were scored as 0. Percentages of mean accuracy and chance-corrected recognition rates were computed. Chance-corrected recognition rates (CCR; Lassalle et al., 2019) were computed instead of unbiased hit rates (H_u; Wagner, 1993; see Chapter 3 for comparison) to retain information about intended intensity. Negative CCR values are reported as 0. CCR was calculated as:

$$CCR = \frac{\frac{Accuracy\ in\ Percent}{100} - \frac{number\ of\ correct\ choice}{number\ of\ all\ choices}}{\frac{number\ of\ incorrect\ choices}{number\ of\ all\ choices}}$$

First mean accuracy and CCRs were calculated for each stimulus (i.e. per speaker ID, intended emotion, intended intensity, and perceived emotion category). Subsequently, the by-item values were used to compute summary statistics (means, sd, range) per intended emotion, intensity, and perceived emotion category. This was done to account for slight variations in listener numbers contributing to the rates of each individual stimulus.

Perceived emotion overall and by intended intensity. Overall emotion accuracy was 38.7% for female and 38.8% for male speaker stimuli (chance-corrected values: 30.9% for female and 31.0% for male stimuli). Table 21 shows Mean recognition accuracy and chance-corrected recognition rates (CCR) per emotion category and intended intensity (see Supplementary material 2 for

values by speaker sex). Stimuli of low intensity seem to be less well recognised than stimuli of high intensity. In both, low- and high-intensity stimuli, there appears to be a large variability within each emotion category.

Table 21: Mean recognition accuracy and chance-corrected recognition rates (CCR) per emotion category and intended intensity

		Mean Recognition Accuracy			Mean Chance-Corrected Recognition Rates (CCR)		
Emotion	Intended Intensity	Mean	SD	Range	Mean	SD	Range
HAP	Low	28.73	25.76	0.00-94.83	21.93	26.04	0.00-94.09
	High	41.59	26.86	0.00-91.94	34.64	28.83	0.00-90.78
SAD	Low	41.32	19.81	8.45-82.09	33.13	22.33	0.00-79.53
	High	42.67	24.11	1.72-82.54	35.90	25.31	0.00-80.05
ANG	Low	28.42	17.85	1.52-77.94	19.79	18.34	0.00-74.79
	High	45.24	22.36	8.62-87.10	37.60	25.25	0.00-85.25
SUR	Low	35.76	18.05	8.62-73.44	26.93	20.13	0.00-69.64
	High	41.52	15.49	9.23-66.67	33.32	17.37	0.00-61.90
FEA	Low	40.06	24.92	3.17-87.93	32.50	27.12	0.00-86.21
	High	61.15	20.69	32.35-93.10	55.60	23.64	22.69-92.12
DIS	Low	17.36	11.69	1.69-51.52	7.65	11.48	0.00-44.59
	High	26.66	15.99	4.69-56.06	17.14	17.15	0.00-49.78
NEU	Normal	53.09	15.86	30.16-90.00	46.39	18.12	20.18-88.57

***Note*. All values in Percent. HAP = Happiness, SAD = Sadness, ANG = Anger, SUR = Surprise, FEA = Fear, DIS = Disgust. SD = Standard Deviation.**

An overview of mean recognition accuracy per emotion category and intensity by speaker sex can be seen in the confusion matrices in Figure 21. Sadness appears to have higher recognition accuracy when enacted by female speakers, but accuracy for the low- and high-intensity conditions were similar within each speaker sex. Happiness, anger, and fear seem to have larger differences between male and female encoders in the high- but not in the low-intensity condition. Happiness appears better recognised when encoded from female speakers, whereas anger, surprise, fear, disgust, and neutral showed marginally higher recognition accuracy from male speakers. Disgust from both speaker sexes was not well recognised in either intensity attempts, yet, it scored (slightly) above the chance-level of 12.5% (from an 8-AFC task).

Figure 21: Confusion matrix of recognition accuracy, separated by speaker sex and intended intensity levels

***Note*. Neutral was neither portrayed as a low- nor high-intensity condition, yet was added to both for comparison. All values reported in %.**

Confusion patterns in Figure 21 appear similar across both speaker sexes. All low-intensity emotions are confused with neutral, however, this is not observable in the high-intensity condition. Disgust is mistaken as either anger or sadness depending on speaker sex and intensity category. Fearful vocalisations in the low- but not high-intensity versions are mistaken as sadness. Intended happiness in the high- but not low-intensity representation is getting confused with surprise, whereas surprise in both high- and low-intensity attempts is being mistaken as happiness.

Recognisability and Authenticity. For each stimulus, recognisability and authenticity were averaged across listeners' 4-point Likert scales responses respectively. An overall composite score (i.e. the average between recognisability and authenticity) was also determined. Intraclass correlation

coefficients (ICC) and their 95% confidence intervals (CI) were computed separately for recognisability and authenticity. Since subsets of listeners rated subsets of vocal stimuli, ICC was based on mean-ratings (with variable k) with absolute agreement using a one-way model (Koo & Li, 2016; Shrout & Fleiss, 1979). According to Koo and Li (2016), the level of reliability can be interpreted as excellent for both recognisability (ICC = .964; CI = [.958; .969]), and authenticity (ICC = .940; CI = [.930; .949]). These values are in accordance with ICCs and interrater reliability reported elsewhere (Lima et al., 2013; Morningstar, Ly, et al., 2018).

4.5.2.2 Listener sex differences

A binomial mixed-effects model with by-item and by-participant random intercepts (see model (2) below) was computed to determine whether correct response was influenced by listener sex, emotion/intensity, speaker sex, and interactions between the factors were added. Given the complex design of having a single neutral representation with one intensity level that is neither high nor low, and not wanting to lose information on either intended emotion or intended intensity, both concepts were merged into a single variable with 13 levels (e.g. neutral_normal, happy_high, happy_low, etc.) and deviation-coded accordingly. The neutral_normal level was selected as the reference category. Listener sex and speaker sex were also deviation-coded. Both, listener sex and speaker sex had originally been added to the random structure, however, they were removed after both random terms were found to overfit the model (rePCA in the lme4 package; Bates et al., 2015).

(2) Correct ~ Emotion_Intensity * Listener Sex * Speaker Sex + (1 | Listener ID) + (1 | Speaker ID)

The model returned a main effect for emotion_intensity (Wald $X^2(12) = 989.854$, $p < .001$), as well interactions of emotion_intensity with both speaker sex (Wald $X^2(12) = 186.073$, $p < .001$) and listener sex (Wald $X^2(12) = 33.674$, $p < .001$). No main effects of listener sex and speaker sex were found ($p > .05$). Sidak-corrected contrast comparisons of emotion_intensity and speaker sex showed that sad stimuli, in both high- and low-intensity conditions, had a higher chance of being recognised correctly when encoded by female compared to male

speakers (sad_low: estimate = 0.187, SE = 0.043, z = 4.355, p < .001; sad_high: estimate = 0.195, SE = 0.043, z = 4.500, p < .001). No further contrast comparisons were significant. The interaction of emotion_intensity and listener sex was driven by the main effect of emotion_intensity. Sidak-corrected contrast comparison showed no significant differences between female and male listeners for any emotion_intensity category (all p > .05).

Analysing the main effect of emotion_intensity showed the model-based predicted probability to recognise neutral stimuli was significantly different compared to all 12 other emotion_intensity categories (all p < .002), and high-intensity stimuli had a higher probability of getting recognised than low-intensity stimuli for happiness (estimate = -0.133, SE = 0.018, z = -7.507, p < .001), anger (estimate = -0.175, SE = 0.018, z = -9.759, p < .001), fear (estimate = -0.224, SE = 0.018, z = -12.269, p < .001), and disgust (estimate = -0.095, SE = 0.015, z = -6.169, p < .001). Predicted probability of low- and high-intensity did not differ significantly for sadness (estimate = -0.021, SE = 0.019, z = -1.153, p = .995) and surprise (estimate = -0.055, SE = 0.018, z = -3.038, p = .111). Additionally, there is more variability in the data and therefore more significant contrast comparisons within low- than high-intensity stimuli (see Supplementary material 3).

Given that listener sex differences were non-significant, perceived emotion accuracy and chance-corrected accuracy will be reported in the database irrespective of listener sex.

Recognisability and authenticity. A cumulative link mixed model with by-subject and by-item random slopes and intercepts was computed to investigate potential listener sex differences for recognisability (model (3) below) and authenticity (model (4) below). Again, speaker sex and the interaction between listener sex and speaker sex were added to the model. Listener sex and speaker sex were deviation-coded.

(3) Recognisability ~ Listener Sex * Speaker Sex + (1 + Speaker Sex | Listener ID) + (1+ Listener Sex | Speaker ID)

(4) Authenticity ~ Listener Sex * Speaker Sex + (1 + Speaker Sex | Listener ID) +
(1+ Listener Sex | Speaker ID)

There were no significant main effects of listener sex or speaker sex, or any interaction between the two variables for either the recognisability or the authenticity model (all $p > .05$). The database will therefore report mean recognisability, mean authenticity, and a composite score for each of the 312 stimuli irrespective of listener sex.

4.5.3 Discussion

4.5.3.1 Overall accuracy of emotion ratings

The overall accuracy of 38.7% from the emotion recognition ratings were lower in comparison to the majority of existing databases (e.g. Belin et al., 2008; Castro & Lima, 2010; Lima et al., 2013), yet similar to Laukka et al. (2013) who reported 39% overall accuracy for positive and 45% for negative representations. However, Laukka et al. included a mix of basic and complex emotion categories and presented 8 positive (study 1) and 8 negative (study 2) whereas in this study only 6 basic emotion categories were rated. Whilst our overall chance-corrected recognition rates of 31% were slightly below the 39% reported by Lassalle et al. (2019), individual values for happiness, sadness, and anger were comparable between the two studies. Neutral (difference of 13.4%) and fearful expressions (differences in low: 9.5%; high: 26.6%) achieved higher recognition rates in our study, whereas disgust (differences in low: 50.4%; high: 27.9%) and surprise differences in low: 36.1%; high: 12.7%) were better recognised in the EU-Emotion Voice Database (Lassalle et al., 2019).

One reason for the overall lower recognition rates in this study compared to the majority of literature could be stimulus type. Studies using non-verbal vocalisations (e.g. Belin et al., 2008; Lima et al., 2013) report higher recognition accuracy than those including speech stimuli (e.g. Castro & Lima, 2010; Lassalle et al., 2019). In a young adult listener sample, Hawk et al. (2009) directly compared affect portrayal from non-verbal vocalisations, a short phrase, and face stimuli, and found that whilst affect bursts and face stimuli were recognised similarly across emotion dimensions, they both achieved higher

overall recognition accuracy compared to speech stimuli. However, there is also some variability in findings when using speech stimuli, depending on whether sentences or words are used. Lausen and colleagues (Lausen & Hammerschmidt, 2020; Lausen & Schacht, 2018) included word and sentence stimuli, and reported emotion recognition accuracy from word stimuli as either slightly lower than or on par to sentences. However, values varied depending on emotion category. The word stimuli were either of positive, negative, or neutral connotation but not socially-relevant like the stimuli used in this study. Despite, these findings suggest that distinct emotion categories are "easier" to identify from non-verbal vocalisation and affect burst than speech segments which may explain the discrepancy between the accuracy rates reported here and the majority of published studies.

There was also variability in accuracy between specific emotion categories. It is worth acknowledging that disgusted expressions in this validation study were not very well recognised from either speaker sex. Whilst recognition accuracy surpassed chance levels, values are remarkably low, especially in the low-intensity condition. The low recognition accuracy can be partially explained by stimulus choice. Research has shown disgust to be one of the best identified emotion categories from non-verbal vocalisations and affect bursts (e.g. Belin et al., 2008; Hawk et al., 2009; Laukka et al., 2013; Lausen & Schacht, 2018; Lima et al., 2013; Sauter, Eisner, Calder, & Scott, 2010). Yet, when judged from speech (like in this study), disgust has one of the lowest recognition rates (Banse & Scherer, 1996; Hawk et al., 2009; Lausen & Hammerschmidt, 2020; Lausen & Schacht, 2018). Despite using sentence stimuli and more emotion categories than this study, Lassalle et al. (2019) reported higher recognition accuracy for disgust compared to our results. Whilst this may hint at a perceptual differences within different types of speech stimuli, the results may be due to the inclusion of semantically congruent and incongruent items in the EU-Emotion Voice database (Lassalle et al., 2019). As previous research has shown higher accuracy for stimuli matching in content and prosody (e.g. Min & Schirmer, 2011), this may indicate that disgust from speech may need para- and extra-linguistic cues to be recognised accurately (Laver, 1994). The low recognition accuracy of disgust could also be partially responsible for the lower overall recognition rates in our study. Despite our accuracy for disgust being lower than the one reported

in Lassalle et al. (2019), our results support the trend in the literature that disgust attains low recognition accuracy when judged from speech stimuli.

We also observed a large variability within each emotion category. Recognition rates ranged from around 0% to as high as 90% but the ranges fluctuated depending on speaker sex, emotion category, and intensity levels. This heterogeneity between stimuli was also observed by others (e.g. Chronaki et al., 2015; Lassalle et al., 2019; Zupan, 2015) showing that emotion recognition from vocal stimuli alone is difficult. Yet, high variability is a representation of real life. It should be highlighted though, that listeners' perception was compared to intended emotion to calculate the recognition accuracy, however some of the individual stimuli with very low recognition accuracy were perceived by the majority of listeners as a different emotion category. We still believe this adds value and a unique opportunity of investigating the relationship between intentions and perceptions in the future.

Finally, there is a need to emphasize differences between high- and low-intensity stimuli. Lassalle et al. (2019) indicated emotion categories of disgust and surprise to be recognised better from low-intensity stimuli. Contrastingly, our results showed all low-intensity representations were recognised less accurately than their high-intensity counterparts, although values for sadness were almost on par. We do agree with the explanation Lassalle et al. (2019) provided for the non-significant differences of low- vs high-intensity sadness. They reasoned that some emotion categories (such as sadness) are of low-intensity by nature and therefore listeners are more attuned to encounter that emotion in low-intensity representations in real life. These potential context effects could also account for the low recognition accuracy of disgust, as it is very unlikely to choose a welcoming word (such as "hello") towards a person or an item when feeling disgusted by them. Linguistically, disgust contains vowel sounds such as [ɯ, u, ʌ, ɜ] and fricative [x, ɸ, h] or bilabial nasal [m] consonants (Goddard, 2014) which the word "hello" did not. Future research may want to employ word stimuli with vowel and consonant sounds related to representations of disgust to investigate the low recognition rates of disgust in speech. Overall, including low-intensity emotion categories with low recognition scores, may have contributed to our overall lower accuracy compared to other databases.

4.5.3.2 Absence of listener sex differences

Our results show that female and male listeners' emotion perceptions did not significantly differ, however, this only aligns to some degree with the mixed findings from existing studies. For emotion recognition accuracy, some literature report an overall female advantage decoding emotion correctly (e.g. Belin et al., 2008; Grosbras et al., 2018; Lausen & Schacht, 2018) or for specific emotion categories (Sen et al., 2018), whereas others see encoding abilities of female and male listeners on par (e.g. Amorim et al., 2021; Hawk et al., 2009; Sauter et al., 2013). In this database, combining emotion and intensity into one variable to avoid losing information of the intensity dimension, we report an emotion/intensity by listener sex interaction. However, post-hoc comparisons showed that the interaction was driven by the main effect of emotion/intensity rather than a perception difference between female and male listeners. Therefore, our results are overall in accordance with the literature reporting no significant listener sex differences. Therefore all validation values in the Glasgow vocal emotion and personality corpus are reported irrespective of listener sex.

4.6 Study 3: Vocal stimuli validation on valence and arousal

4.6.1 Methods

4.6.1.1 Power analysis

Similar to studies 1 and 2, PANGEA was used for power calculation. Effect sizes were set to the recommended effect size of 0.45 by Jake Westfall (2016), since no research currently exists investigating listener sex differences of valence or arousal to word stimuli. A minimum of 18 participants per listener sex would be required to achieve power of .9. The study is therefore sufficiently powered.

4.6.1.2 Listeners

Twenty-five female (mean age = 26.96 ± 5.07 years, range: 19-37) and 25 male listeners (mean age = 28.60 ± 7.15 years, range: 18-39) were recruited via Prolific (www.prolific.co; Palan & Schitter, 2018) to take part in the valence and arousal experiment. Eighteen to 39-year-olds were eligible to partake if they had

no hearing deficits, were not currently taking any mental health medication and had a Prolific approval rate of at least 95%. Participants were either currently residing in Scotland, had lived in Scotland at some point in their lives, or had other ties to Scotland (i.e. family/friends). Data from three initial participants were replaced: two females had failed more than 50% of the attention checks, and visual inspection identified inconsistencies of scale use between the 8 blocks for one male listener. The experiment on valence and arousal took approximately 40 min to complete and participants were reimbursed with £5.

Due to an unnoticed coding mistake when analysing attention checks, more than half of the participants appeared to have failed the attention checks on the arousal scale whilst succeeding in at least 87.5% of the 8 valence dimensions. This discrepancy led to 50 new participants being recruited via Prolific (Palan & Schitter, 2018) for an arousal-only experiment (see Table 22 for more detail). Recruitment criteria were as listed above. Data from two initial listeners were replaced due to them either rating 66.3% of the stimuli within a 10% margin of the lower slider extreme (1 female) or showing anomalous rating patterns between the female and male stimuli blocks (1 male; identified via visual inspection). Average completion time for the arousal-only experiment was ca 25 min with monetary incentives of £3 provided.

Table 22: Demographic information of participants in the arousal-only experiment

Listener Sex	N	Mean Age	SD Age	Min Age	Max Age
Female Listeners	25	27.60	5.16	18	39
Male Listeners	25	28.84	6.45	19	39

4.6.1.3 Materials, experimental set-up, and rating procedures

Experiments were created with the PsychoPy Builder interface (v2021.1.4; Peirce et al., 2019) and output to a PsychoJS experiment hosted on Pavlovia (Bridges et al., 2020). Recruitment service Prolific (Palan & Schitter, 2018) was used to advertise the experiment and recruit participants. On Prolific, participants received detailed information about the purpose of the study and instructions about experimental set-up. On Pavlovia, participants provided demographic information before reading brief general information about the

study purpose, providing consent by agreeing to start the experiment, and receiving detailed task instructions. The experiment started with four practice trials to provide an opportunity to familiarise participants with the rating procedures and adjust the volume if necessary.

Within each of the two experiments, listeners would rate all 312 stimuli. The order as to whether the 156 female or male emotive representations were shown first, was counterbalanced. To provide listeners with an option to take small breaks, stimuli were presented in four blocks of 39 stimuli per speaker sex which were created following the rules outlined above in Study 2. The order of stimuli within each block was randomised.

For the valence and arousal experiment, participants would see two questions presented above a VAS slider for each dimension. The question on the valence domain asked “Valence: How positive or negative does this emotion sound to you?” with the VAS slider ranging from “negative” (left) to “positive” (right). For arousal, the question read “Arousal: How does this emotion make you feel?” with slider anchors being “calm/dull” on the left and “excited/agitated” on the right. Stimuli and block counters were added to the bottom of the rating pages to manage participants expectations, placed next to a “next” button. The “next” button could only be pressed after slider responses were made on both rating scales.

The rating procedure for the arousal-only experiment was modified so that only one VAS slider was presented in the middle of the screen with an accompanying question placed above it. The arousal-only question and slider extremes matched the ones from the arousal dimension of the valence and arousal experiment. Similar to the valence and arousal experiment, stimuli and block counters were incorporated again, and a slider response needed to be detected before the “next” button would allow participants to progress.

Attention checks were placed at the end of each of the 8 blocks. Slider positions and anchor labels were presented exactly as in the experimental trials but with instructions to rate the stimulus heard at either extreme end of a particular slider. Attention checks for valence and arousal included two sliders with instructions above each to “Rate this emotion as [*negative* or *positive*]” on the

valence slider and “Rate this emotion as [*calm*/*dull* or *excited*/*agitated*]” on the arousal slider. The arousal-only experiment included a single slider with the same [*calm*/*dull* or *excited*/*agitated*] endpoints, however, instructions were slightly modified in comparison to the valence and arousal experiment to account for the assumed rating difficulties (i.e. “This emotional expression is making you feel [*calm*/*dull* or *excited*/*agitated*]. Rate it as [*calm*/*dull* or *excited*/*agitated*]).

4.6.2 Results

4.6.2.1 Initial data preparation and reliability

Slider values for valence and arousal were operationalised on a scale from 1 to 100. Attention check questions were tailored to elicit a response towards the ends of the slider extremes. An attention checks counted as “pass” when sliders were within a 20% margin of the requested response. Given attention checks were placed after each experimental block, there were a total of 8. For the valence and arousal study, participants had to answer both questions correctly for the check to count as “pass”. If one of the questions was answered incorrectly, then the attention check would be counted as a “fail”. Cumulatively, participants had to succeed in at least 4 checks for their data to be included in the analysis.

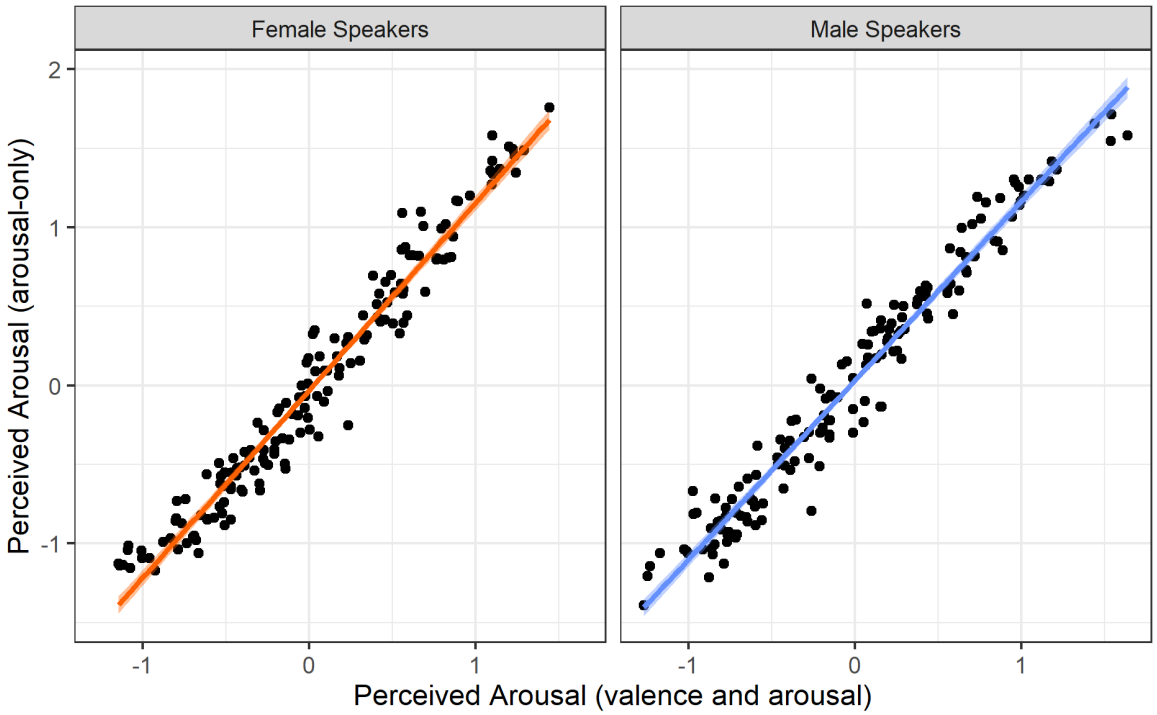
Each participant’s ratings were z scored across all ratings provided, and subsequently averaged for each vocal stimulus to compute mean score for perceived valence, arousal, and arousal (from arousal-only experiment). Z-scores were used to account for potential differences in participants’ use of rating scales. Intraclass correlation coefficients (ICCs) and their 95% CIs were computed on the z-scored values for valence, arousal (from valence and arousal experiment), and arousal (from arousal-only experiment) respectively. Since all stimuli were rated by all listeners, who are the only raters of interest, ICC was based on a two-way mixed effects model using the mean ratings of k raters’ consistency as a basis (Koo & Li, 2016; Shrout & Fleiss, 1979). Inter-rater consistency is considered excellent for all perceived emotion dimensions (Koo & Li, 2016) and is similar to interrater reliability from other corpora (e.g. Lima et al., 2013). See Table 23 below for details.

Table 23: Intraclass correlation coefficients (ICC) with 95% confidence intervals (CI) for each of the perceived emotion dimensions valence, arousal, and arousal (arousal-only)

Perceived Emotion Dimension	ICC	CI Lower Boundary	CI Upper Boundary
Valence	.975	.971	.979
Arousal	.972	.968	.976
Arousal (arousal-only)	.986	.984	.988

To investigate the relationship between the arousal ratings obtained in the valence and arousal experiment and the arousal-only dataset, Pearson correlation coefficients were computed on the z-scored data. Results returned very strong positive relationships for the female (r = .979) as well as the male speaker sex (r = .980; see Figure 22). This shows that participants were able to provide valid ratings for the arousal dimension, even when assessed together with valence.

Figure 22: Scatterplot of z-scored arousal ratings from the valence and arousal and the arousal-only experiment



Note. **Each dot represents a vocal stimulus. Line of best fit was added.**

4.6.2.2 Listener sex differences

Linear mixed-effects models with by-subject random intercepts and slopes and by-item random intercepts were computed to investigate potential listener sex differences for the valence (see model (5) below) and arousal data (see model

(6) below). Speaker sex and the listener sex by speaker sex interaction were included as co-variates, and listener sex and speaker sex were deviation-coded. Listener sex was originally included in the by-item random effect structure of each model, yet was removed after the PCA analysis of random-effects variance-covariance estimates showed overfitting (rePCA in the lme4 package; Bates et al., 2015).

(5) Valence ~ Listener Sex * Speaker Sex + (1 + Speaker Sex | Listener ID) +
(1 | Speaker ID)

(6) Arousal ~ Listener Sex * Speaker Sex + (1 + Speaker Sex | Listener ID) +
(1| Speaker ID)

The valence model revealed no main effect of listener sex, speaker sex, or the interaction between both. The arousal model, however, showed a significant main effect of listener sex ($X^2(1) = 5.956$, $p = .015$) which was further quantified by a significant interaction of listener and speaker sex ($X^2(1) = 4.288$, $p = .038$). Sidak-corrected contrast comparisons (package emmeans; Lenth, 2021) showed that female listeners were predicted to rate male but not female speakers' affective vocalisations as more arousing than male listeners (estimate = 6.55, SE = 2.16, z = 3.028, p = .010).

We further investigated whether the effect of listener sex held true for the arousal-only experiment. The model deviated slightly from model (6) in that a random slope for listener sex was added to the by-item random effect structure. However, all other aspects were kept identical (see model (7) below).

(7) Arousal ~ Listener Sex * Speaker Sex + (1 + Speaker Sex | Listener ID) +
(1 + Listener Sex| Speaker ID)

Results showed no significant main effects of listener or speaker sex, or the interaction term (all sidak-corrected $p > .05$). We therefore combined the arousal data from the valence and arousal experiment with the data from the arousal-only experiment. rePCA (Bates et al., 2015) showed overfitting for the by-item random effect structure when listener sex was included, so we computed model (6) again. Results revealed no main effect of listener sex,

speaker sex or interaction of listener and speaker sex (all $p > .05$)[8]. We conclude that significant results for arousal obtained in the analysis of the valence and arousal data were due to chance. Therefore, the database will host values for valence, arousal, arousal-only, and a combined arousal score regardless of listener sex.

4.6.3 Discussion

We did not detect listener sex differences for valence and arousal. To our knowledge, there are not many studies investigating listener sex differences on vocal dimensions of valence and arousal. Nevertheless, our results are in agreement with results published by Belin et al. (2008) and Koeda et al. (2013). Despite listeners assessing vocal stimuli either from speakers of the same (this study; Belin et al., 2008) or different cultural background (Koeda et al., 2013), there was no association between listener gender and perceived valence and arousal. This might hint at universality for valence and arousal but needs to be investigated further in future.

4.7 Study 4: Vocal stimuli validation on perceived emotional intensity

4.7.1 Methods

4.7.1.1 Power analysis

Power analysis was conducted via PANGEA. Similar to Study 3, the authors are not aware of any research investigating listener sex differences of perceived emotional intensity. Therefore, we used the recommended effect size of 0.45 (Westfall, 2016) to compute the minimum number of listeners to participate to achieve power of .9. Equivalent to Studies 1 and 3, the minimum of listeners needed to be recruited is 18. The study is therefore sufficiently powered.

[8] Model 7, including listener sex in the random-effect structure, was additionally computed on the arousal data (from the valence and arousal experiment), and the composite data from the arousal and arousal-only experiment to test whether the modification of the random structure could explain the difference in results. Whilst Chi-square estimates and p-values differed slightly, the overall results were not affected by including listener sex in the by-item random effect structure.

4.7.1.2 Listeners

In total, 25 female (mean age = 27.92 ± 6.22 years, range: 18-39) and 25 male (mean age = 30.12 ± 6.41 years, range: 18-39) participants were recruited via Prolific (www.prolific.co; Palan & Schitter, 2018) to take part in the perceived intensity experiment. Recruitment criteria are equivalent to the ones outlined in Study 3. Data from two male participants were replaced due to one failing more than 50% of the attention checks and the other having one aborted and one completed experiment attempt with different scale use patterns. Completion time was approximately 25 min and participants were reimbursed with £3.

4.7.1.3 Materials, experimental set-up, and rating procedures

The experiment for perceived intensity was created, hosted, and advertised equivalent to Study 3. Participants would encounter the same instructions and rating procedures as outlined in the arousal-only experiment in Study 3, being presented with a single VAS slider. Participants were asked to rate "Intensity: How intensely is this emotion expressed by the speaker?" on a slider ranging from "very subtle" (left) to "very intense" (right). Again, stimuli and block counters were shown on the page, and participants would only be allowed to progress to the next trial once they selected a slider response. Attention checks for perceived intensity included one slider and instructions asked participants to "Rate this emotion as [*very subtle* or *very intense*]".

4.7.2 Results

4.7.2.1 Initial data preparation and reliability

The procedures to analyse attention checks, perform z-scoring, and compute Intraclass correlation coefficients were exactly the same as outlined in section 4.6.2.1 of Study 3. Inter-rater consistency (ICC = .981, CI = [.978; .984]) is considered excellent for perceived intensity (Koo & Li, 2016).

4.7.2.2 Listener sex differences

A linear mixed-effects model with by-subject and by-item random intercepts and slopes was computed to predict perceived intensity from listener sex (see model

(8) below). Speaker sex and the listener sex by speaker sex interaction were included as control variable. Listener sex and speaker sex were deviation-coded.

(8) Intensity ~ Listener Sex * Speaker Sex + (1 + Speaker Sex | Listener ID) + (1 + Listener Sex| Speaker ID)

No significant main effects of listener or speaker sex, or interaction between the two variables were found (all $p > .05$). Therefore, the values of perceived intensity are reported within the database irrespective of listener sex.

4.7.3 Discussion

In this study, female and male listeners formed similar perceptions about the intensity of the portrayed vocal emotions. This is only in partial agreement with the existing literature. Koeda et al. (2013) found no decoder sex differences of intensity when presenting the MAV stimuli to Japanese listeners. However, Belin et al. (2008) reported that male listeners provided slightly higher intensity ratings compared to female participants. This may indicate cross-cultural differences. Both studies used the MAV stimuli (Belin et al., 2008) encoded by French-Canadian speakers, however listener groups were either the same culture as the speakers (French-Canadian listeners; Belin et al., 2008) or different (Japanese listeners; Koeda et al., 2013). Japanese listeners as an out-group may not have been able to detect fine-tuned differences within the expressions compared to in-group French-Canadian listeners (Koeda et al., 2013). Yet, in this study, there were no significant differences between female and male listeners' perception of intensity, despite all listeners being familiar with Scottish voices. It may be speculated that differences arise due to stimulus choice since the MAV includes affect bursts and not words of socially-relevant content, however, this is a research field that would require further investigation.

4.8 General discussion

The Glasgow vocal emotion and personality corpus is a collection of 312 affect stimuli validated on a variety of personality traits and affect measures. Some databases (e.g. Belin et al., 2008; Lassalle et al., 2019; Lima et al., 2013) included validated scores of valence and arousal with or without intensity in

addition to emotion categories. Lima and colleagues added recognisability (labelled as intensity; Castro & Lima, 2010) and authenticity (Lima et al., 2013) to categorical emotions. However, we are not aware of any database including valence, arousal, intensity, recognisability and authenticity, as well as validation scores on trustworthiness, dominance, and attractiveness. Despite further analysis being outwith the scope of this thesis, we also collected free-response ratings for the 312 stimuli and extracted acoustic values. The affect representations of the word "hello" were chosen as a starting point for this database trying to close the gap in the literature by adding socially-relevant word stimuli. Furthermore, we are contributing to the open-science movement by making this extensive dataset freely and openly available (Creative Commons Attribution 4.0 International License). This rich corpus would not only be useful for researchers in the field of emotion and/or personality, but also in a teaching setting.

Overall, this study has shown that humans make very quick judgements about vocal affect representations of socially-relevant stimuli that are significantly better than chance. Given that there are no direct comparisons in the open-access speech database literature to compare our results to, our findings suggest that brief socially-relevant words appear more closely related to speech stimuli than non-verbal vocalisations since overall accuracy is similar to findings reported in other databases and papers using speech stimuli. However, there is substantial variability within and between emotion categories which may be a reflection of real life. Hence, this validation study has shown that the stimuli presented in the Glasgow vocal emotion and personality corpus are representative and contain ecological validity.

However, this validation study is not without limitations. One caveat of this study could be that we did not provide speakers with vignettes for each of the emotion categories they were about to encode. The majority of researchers tend to provide scenarios/ instructions to guide speakers towards a specific emotion, provide feedback or ask them to produce affective vocalisation until the researcher is satisfied those affect portrayals are recognisable (Bänziger et al., 2012; Belin et al., 2008; Burkhardt et al., 2005; Castro & Lima, 2010; Laukka et al., 2010; Lima et al., 2013; Sauter, Eisner, Calder, & Scott, 2010). In our

opinion, not all speakers may feel equally emotive to provided scenarios, and satisfying the researcher's standards of how an emotion should be portrayed, could result in overly stereotypical vocalisations. Here, we deliberately chose not to provide specific vignettes or feedback for speakers to elicited emotions to be representative of the encoder. By creating a corpus of natural-sounding yet acted vocalisations, we may have sacrificed recognition rates, but we believe they are more representative of real-life scenarios.

Our future aim is to gradually expand this database by adding further validated stimuli types from the same speakers, such as affect bursts, vowels, emotive words and pseudo-sentences, and neutral reading scenarios. This would provide us and other researchers with an opportunity to not only investigate emotion or personality perceptions from a variety of stimuli types but also study the relationship between the two concepts in depth in the future. Additionally, we aim to include the speakers' free speech and creative speech stimuli as examples of natural speech and investigate similarities and differences between natural speech and read out or enacted scenarios. Another research avenue to pursue further is the acoustic analyses on these vocalisations to investigate which specific acoustic measures contribute to perceptions of emotion and personality.

4.9 Summary of the measures reported in the open-access database

To summarise, the Glasgow vocal emotion and personality corpus is an open-access database that will hold the following baseline measures for each of the 312 vocal stimuli:

- The actor's intended emotion and their intended intensity (as noted during the recording procedure);
- Ratings of perceived trustworthiness, dominance, and attractiveness;
- Recognition accuracy and chance-corrected recognition rates;
- Recognisability, authenticity, and a composite score of both;

- Valence and arousal ratings;
- Perceived intensity;
- Categorical labels from a free-labelling task[9];
- Acoustic measures[10].

[9] The data from the free-labelling tasks are still being analysed. Therefore, Study 5 is excluded in the thesis.

[10] It should be noted, that the collection of acoustic measures was addressed in section 4.2.6 but not analysed further.

4.10 Supplementary material 1

Rainbow Passage (abridged)

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colours. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

Throughout the centuries people have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods.

Telephone Scenario 1

Emma is reading a book in her room. Suddenly, the phone rings, “Hello?!” she answers in a curious voice. What a surprise to hear the voice of her friend who had been gone on a road trip for the last 4 months.

Telephone Scenario 2

As I sat in my room, suddenly the phone rang. The voice said: ‘Hello. This is your lecturer. I urge you to submit your essay by the end of the week.’ What a surprise it was to receive such a call on a Sunday night!

Harvard Sentences

H1 Harvard Sentences	H2 Harvard Sentences
1. The birch canoe slid on the smooth planks.	1. The boy was there when the sun rose.
2. Glue the sheet to the dark blue background.	2. A rod is used to catch pink salmon.
3. It's easy to tell the depth of a well.	3. The source of the huge river is the clear spring.
4. These days a chicken leg is a rare dish.	4. Kick the ball straight and follow through.
5. Rice is often served in round bowls.	5. Help the woman get back to her feet.
6. The juice of lemons makes fine punch.	6. A pot of tea helps to pass the evening.
7. The box was thrown beside the parked truck.	7. Smoky fires lack flame and heat.
8. The hogs were fed chopped corn and garbage.	8. The soft cushion broke the man's fall.
9. Four hours of steady work faced us.	9. The salt breeze came across from the sea.
10. A large size in stockings is hard to sell.	10. The girl at the booth sold fifty bonds.

Word list

- North
- White
- Echo
- South
- Green
- Light
- Delta
- Up
- Red
- Right
- Start
- Left
- Bottom
- Stop
- Yellow
- Top
- Blue
- East
- Down
- Bravo
- West
- Charlie
- Bright
- Alpha
- Good-bye
- Night

Affective Reading – Pseudo-Sentences

- I tropped for swinty gowers.
- She kuvelled the noralind.
- The placter jabored the tozz.
- The moger is chalestic.
- The rivix joled the silling.
- The crinklet is boritate.
- She krayed a jad ralition.
- We wanced on the nonitor.
- They pannifered the moser.
- We groffed for vappy laurits.
- I marlipped the tovity.
- The varmalit was raffid.
- They rilted the prubition.
- Ne kalibam sout molem.

4.11 Supplementary material 2

Table 24: Mean recognition accuracy and chance-corrected recognition rates (CCR) per emotion category and intended intensity, separately by speaker sex

		Mean Recognition Accuracy			Mean Chance-Corrected Recognition Rates (CCR)		
Emotion	Intended Intensity	Mean	SD	Range	Mean	SD	Range
				Female Speakers			
HAP	Low	30.43	29.10	0.00-94.83	24.01	29.77	0.00-94.09
	High	46.81	31.09	0.00-91.94	40.90	33.06	0.00-90.78
SAD	Low	50.12	20.37	20.00-82.09	43.00	23.28	8.57-79.53
	High	52.27	22.89	13.64-82.54	45.45	26.16	1.30-80.05
ANG	Low	26.06	12.04	4.84-40.91	16.54	12.04	0.00-32.47
	High	41.68	23.91	8.62-87.10	33.71	26.80	0.00-85.25
SUR	Low	33.78	19.13	8.62-73.44	25.01	20.94	0.00-69.64
	High	38.95	18.02	9.23-66.67	30.54	20.06	0.00-61.90
FEA	Low	37.55	23.58	3.17-75.76	29.75	25.39	0.00-72.29
	High	55.40	17.30	32.81-84.85	49.03	19.77	23.21-82.68
DIS	Low	14.42	9.11	1.69-34.85	5.12	7.36	0.00-25.54
	High	25.93	15.60	7.46-56.06	16.12	16.96	0.00-49.78
NEU	Normal	49.66	13.96	30.16-70.97	42.47	15.96	20.18-66.82
				Male Speakers			
HAP	Low	27.02	23.11	0.00-71.19	19.84	22.85	0.00-67.07
	High	36.37	21.97	8.45-67.61	28.37	23.64	0.00-62.98
SAD	Low	32.51	15.38	8.45-62.07	23.25	16.96	0.00-56.65
	High	33.08	22.15	1.72-62.07	26.34	21.35	0.00-56.65
ANG	Low	30.77	22.56	1.52-77.94	23.05	23.14	0.00-74.79
	High	48.81	21.11	16.90-81.82	41.49	24.12	5.03-79.22
SUR	Low	37.74	17.52	16.67-66.18	28.84	20.03	4.76-61.34
	High	44.09	12.74	28.77-64.62	36.10	14.56	18.59-59.56
FEA	Low	42.58	27.00	4.11-87.93	35.25	29.61	0.00-86.21
	High	66.89	22.88	32.35-93.10	62.16	26.15	22.69-92.12
DIS	Low	20.29	13.56	7.58-51.52	10.18	14.40	0.00-44.59
	High	27.39	17.02	4.69-55.88	18.17	18.02	0.00-49.58
NEU	Normal	56.53	17.46	33.33-90.00	50.31	19.96	23.81-88.57

***Note*. All values in Percent. HAP = Happiness, SAD = Sadness, ANG = Anger, SUR = Surprise, FEA = Fear, DIS = Disgust. SD = Standard Deviation.**

4.12 Supplementary material 3

Contrast comparisons between emotion categories for low- (Table 25), and high-intensity (Table 26) stimuli.

Table 25: Contrast comparisons for model-based predicted probability between emotion categories for low-intensity stimuli

Emotion Contrasts	Estimate	Standard Error	Z ratio	P value
Happiness - Sadness	-0.119	0.018	-6.680	<.001
Happiness - Anger	0.003	0.017	0.161	1.000
Happiness - Surprise	-0.072	0.017	-4.203	0.002
Happiness - Fear	-0.116	0.018	-6.580	<.001
Happiness - Disgust	0.117	0.016	7.355	<.001
Sadness - Anger	0.122	0.018	6.826	<.001
Sadness - Surprise	0.047	0.018	2.560	0.335
Sadness - Fear	0.003	0.018	0.172	1.000
Sadness - Disgust	0.235	0.018	13.023	<.001
Anger - Surprise	-0.075	0.017	-4.356	<.001
Anger - Fear	-0.118	0.018	-6.728	<.001
Anger - Disgust	0.114	0.016	7.206	<.001
Surprise - Fear	-0.043	0.018	-2.413	0.434
Surprise - Disgust	0.189	0.017	10.954	<.001
Fear - Disgust	0.232	0.018	13.008	<.001

***Note*. All comparisons are tukey-corrected. The estimate is the difference of predicted probability to respond correctly between the compared emotion categories. Df = inf for asymptotic test.**

Table 26: Contrast comparisons for model-based predicted probability between emotion categories for high-intensity stimuli

Emotion Contrasts	Estimate	Standard Error	Z ratio	P value
Happiness - Sadness	-0.007	0.019	-0.394	1.000
Happiness - Anger	-0.039	0.019	-2.130	0.642
Happiness - Surprise	0.006	0.018	0.307	1.000
Happiness - Fear	-0.206	0.018	-11.263	<.001
Happiness - Disgust	0.155	0.018	8.767	<.001
Sadness - Anger	-0.032	0.019	-1.722	0.886
Sadness - Surprise	0.013	0.019	0.699	1.000
Sadness - Fear	-0.199	0.019	-10.774	<.001
Sadness - Disgust	0.162	0.018	9.086	<.001
Anger - Surprise	0.045	0.019	2.439	0.416
Anger - Fear	-0.167	0.018	-9.052	<.001
Anger - Disgust	0.194	0.018	10.861	<.001
Surprise - Fear	-0.212	0.018	-11.594	<.001
Surprise - Disgust	0.149	0.018	8.471	<.001
Fear - Disgust	0.361	0.018	20.610	<.001

***Note*. All comparisons are tukey-corrected. The estimate is the difference of predicted probability to respond correctly between the compared emotion categories. Df = inf for asymptotic test.**

Chapter 5 General Discussion

This thesis had three overarching aims: Firstly, it targeted to close the gap in the current literature as to how perceptions of vocal trustworthiness develop and mature across the early lifespan (Chapter 2). Secondly, it aimed to expand the current literature regarding the early developmental trajectory of perceived vocal emotion by employing socially-relevant stimuli (Chapter 3). Thirdly, this thesis intended the creation of an open-access vocal database validated on a variety of personality and emotion dimensions (Chapter 4). This database would not only be valuable for researchers to study perceptions of vocal personality and/or vocal emotion, but may also be used in a teaching setting.

Section 5.1 will summarise the key findings from each of the experimental chapters, before highlighting implications (section 5.2), and addressing limitations and future directions (section 5.3).

5.1 Summary of main findings

Chapter 2 determined the developmental course of perceived vocal trustworthiness between childhood and early adulthood applying linear mixed-effects modelling. We found a small but statistically significant increase of perceived trustworthiness with listener age, showing our perceptions become slightly more positive as we age. However, we also found very strong correlations of vocal trustworthiness perceptions between age groups when dividing participants into chronological age categories of children (5-10 years), adolescents (11-19 years), and young adults (20-29 years). The correlation analysis showed that children are able to assess not only who sounds trustworthy or not, but are able to judge on a continuum of trustworthiness. Furthermore, it was established that results were independent of listener sex implying that trustworthiness perceptions develop similarly between female and male listeners. Exploratory analysis revealed the difference between 1st and 2nd rating and the range of scale use decreased significantly with an increase in listener age. Taken together, the results from chapter 2 suggest that an understanding of vocal trustworthiness already exists by 5 years of age but perceptions become more “fine-tuned” and “nuanced” with increasing age.

Chapter 3 of this thesis aimed to investigate the developmental trajectories of vocal emotion recognition between childhood and early adulthood (5 to 39 years of age). We found that children are able to recognise vocal emotion at higher than chance levels, however, that ability improved significantly with increasing age. The results from the age group analysis imply that overall accuracy improved significantly between childhood and adolescence and very little thereafter. It was shown that surprise and disgust were recognised with significantly lower accuracy than happiness, sadness, anger, fear, and the neutral representation. Furthermore, findings suggested that the developmental course of emotion across age was different for individual emotion categories. Steeper development was observed for sadness, fear and neutral, whereas happiness, anger, and surprise matured more gradually. Whilst surprise recognition improved with increasing age, disgust remained a poorly recognised emotion category across age.

We also investigated whether females and males differed significantly either in decoding or encoding emotion, and found no significant main effect of either listener sex or speaker sex. However, both variables had significant interactions with emotion. Females were better at recognising disgust, whereas males achieved higher recognition rates identifying neutral expressions. Whilst the listener sex by emotion interaction was statistically significant, we question whether this result is a meaningful one given the overall low recognition rate of disgust. Therefore, we draw attention to the non-significant differences for happiness, sadness, anger, surprise, and fear, and further suggest that results should be replicated before drawing concrete conclusions. In regard to speaker sex, males were better encoders for angry expressions. Again, this result warrants replication.

In **Chapter 4**, we created and validated a vocal emotion and personality database with stimuli from the same speakers. In addition to expanding on existing corpora by including socially-relevant stimuli, we also validated them on a variety of dimensions, such as perceived personality traits (i.e. trustworthiness, dominance, and attractiveness), perceived emotion category, recognisability and authenticity, valence and arousal, perceived intensity, and acoustic measures. Free response ratings will be added in the near future.

Therefore, this rich database not only allows researchers of vocal personality and emotion to study these concepts with stimuli of high ecological validity, but it also paves the way to study the concepts in relation to one another in the future. By making this rich database openly available and free of charge (Creative Commons Attribution 4.0 International License), it is a valuable contribution to open science.

Table 27 summarises the most important findings from the experimental chapters.

Table 27: Summary of key findings in this thesis

Chapter and Title	Key Findings
Chapter 2 Development of perceived vocal trustworthiness across the early lifespan	• Children are able to form impressions of how trustworthy a speaker sounds • Children perceive vocal trustworthiness on a continuum • Vocal trustworthiness perceptions become slightly more positive between childhood and young adulthood • Perceptions become more "fine-tuned" and "nuanced" with increasing age
Chapter 3 Development of perceived vocal emotion across the early lifespan	• Children can recognise the basic 6 emotion families accurately (at above chance levels) • Vocal emotion recognition improves significantly between childhood and adolescence, with little improvement between adolescence and adulthood • Some emotion categories (sadness, fear, and neutral) improve more rapidly whilst others (happiness, anger, and surprise) mature more gradually • Disgusted expressions are recognised poorly by all listeners
Chapter 4 The Glasgow vocal emotion and personality corpus	• Expands the field by adding socially-relevant affect stimuli from the same speakers, therefore increasing ecological validity • Stimuli are validated on multiple emotion and personality measures • Closes the gap in the literature to include a rich corpus that allows the study of vocal personality and emotion separately and simultaneously in the future • Is a valuable contribution to open science

5.2 Implications

5.2.1 Supporting future research on perceived personality and emotion

Both vocal trustworthiness and emotion judgements are made after brief exposure; both are said to play a role in approach/avoidance. Face research has suggested emotion and personality are linked via the emotion overgeneralisation hypothesis (McArthur & Baron, 1983; Montepare & Dobish, 2003; Zebrowitz & Collins, 1997; Zebrowitz & Montepare, 2008). According to that account, there are subtle affect cues in a facial expression (even in an emotionally neutral one) that serve as a baseline judgement of more stable personality traits. Rather than being just a simple judgement of how angry someone feels, we infer social information, such as how likely someone is to attack us or how unfriendly or domineering they are in general (Montepare & Dobish, 2003).

Similarly, voice research has started to investigate the emotion overgeneralisation hypothesis to better understand how emotion and personality traits are connected to one another in the vocal domain (Berry, 1990; Hall et al., 2017; Pinheiro et al., 2021; Schirmer et al., 2019; Zebrowitz-McArthur & Montepare, 1989). For example, a childlike voice is a predictor of warmth (Zebrowitz-McArthur & Montepare, 1989), and both non-verbal vocalisations, such as laughs and cries (Pinheiro et al., 2021), and perceived speaker valence (Schirmer et al., 2019) have been shown to relate to trustworthiness impressions. These findings seem to suggest that there is some connection between these brief judgements of emotion and personality, though exact mechanisms remain to be investigated.

Given the developmental evidence presented in this thesis, we question whether the emotion generalisation hypothesis can explain the relationship between vocal emotion and personality in full. Findings from Chapter 2 suggested that children were able to form trustworthiness perceptions on a continuum. Despite these impressions becoming slightly more positive, consistent, and more nuanced during maturation, there was only slight improvement to achieve adult-like consensus. Contrastingly, results from Chapter 3 seem to suggest that the ability to recognise vocal emotion accurately improved considerably between

childhood and adolescents, despite children being able to detect emotion at better-than-chance levels. These findings suggest that there is more noticeable maturation on the affect than on the trustworthiness dimension. Therefore, the question emerges whether we are indeed extracting subtle emotion cues to form trait impressions about a speaker. We may only speculate here since the developmental studies did not use the same stimuli, and those used in Chapter 2 did not vary explicitly on emotional prosody.

Future research should investigate how emotion and personality are related in adults and whether/how they may influence one another. It would also be of interest to explore whether children apply different strategies or rely on information other than emotion to base their trustworthiness impressions upon. The Glasgow vocal emotion and personality corpus has been established with the goal to support perception researchers to study the concepts of personality and emotion concurrently.

5.2.2 Early emergence of a positivity effect of trustworthiness

An interesting finding that emerged was the slight positivity effect in vocal trustworthiness (Chapter 2). Usually, an age-related positivity effect is reported to emerge in mid- to late adulthood when people start favouring positive over negative stimuli (Reed & Carstensen, 2012). Carstensen and colleagues (Carstensen, 2006; Carstensen & DeLiema, 2018; Carstensen et al., 1999; Reed & Carstensen, 2012) positioned the positivity effect within the framework of socioemotional selectivity theory (SST) which sees the perception of future time as a core construct in the pursuit of social goals. If time horizon is perceived to be limited, as typically observed in older age, present-oriented goals relating to emotional gratification (such as attending to positively-valenced stimuli) are favoured. Contrastingly, when the time horizon is perceived as non-limited, knowledge-related goals with long-term rewards are favoured.

Whilst an age-related positivity effect has been shown to exist in emotional judgements of speech and music (Laukka & Juslin, 2007; Lima & Castro, 2011; Parks & Clancy Dollinger, 2014; but see Amorim et al., 2021), and has not been studied explicitly in regard to perceived vocal trustworthiness, it was surprising to detect the effect between childhood and young adulthood. This finding may

suggest that the positivity effect is not an age-related one per se, but perhaps a gradual effect spread across the entire lifespan or that it may start much earlier in life than previously assumed. Children, adolescents, and young adults are expected to perceive the future as “long and nebulous” (Reed & Carstensen, 2012, p. 1) if the motivational framework of SST and the limited time-horizon explanation were to apply. Therefore, this suggests that an age-related positivity effect across the early lifespan may have different mechanisms and motivations to be explained in full.

It also needs to be highlighted that the effect we found in Chapter 2 was very small. Yet, sensitivity power analysis suggested the study was sufficiently powered to find this small of an effect. Here we would like to encourage future research to replicate the finding and expand investigation beyond early adulthood to discover whether this gradual positivity trend continues into middle and late adulthood or whether it plateaus past age 29 before re-emerging at a later stage in life. Given the stimuli in Chapter 2 were intended to be emotionally neutral, future research may also want to collect valence perceptions on these stimuli to investigate whether participants extracted underlying affect information from the emotionally-neutral stimuli.

5.2.3 The importance of stimuli selection and analysis methods

We observed slight differences in the listener sex by emotion and the speaker sex by emotion interaction between Chapters 3 and 4. In Chapter 4, we reported an emotion/intensity by listener sex interaction which is similar to the emotion by listener sex interaction reported in Chapter 3. The post-hoc comparisons in Chapter 4 showed that the interaction was driven by the main effect of emotion/intensity. Contrastingly, in Chapter 3, we reported a very small advantage for females recognising disgust, and for males recognising neutral expressions. Additionally, when treating age as a continuous variable in Chapter 3, the emotion by listener sex interaction was non-significant. Similarly, there are discrepancies in the findings between chapters 3 and 4 regarding the interaction of speaker sex and emotion. In Chapter 3, males seem to express anger better than females, whereas in Chapter 4, sad stimuli, had a higher chance of being recognised correctly when encoded by female compared to male speakers. This held true for both high- and low-intensity emotion portrayals.

These differences in listener and speaker sex interactions may depend on the specific stimuli included in a study. In Chapter 4, the full stimuli set (i.e. 312 stimuli) was validated, whereas in Chapter 3, only a small subset of those stimuli was utilised (i.e. 35 stimuli). Given not all stimuli were recognised extremely well in Chapter 3, the pre-selection of stimuli to include in subsets becomes crucial. The variability in the number of stimuli included and/or how well those are recognised may be a reason for the variability in results relating to listener sex differences in the existing literature. Future research may want to investigate those relationships methodologically in more detail.

Another explanation for the differing results could lie in the analysis method chosen. In Chapter 4, emotion and intensity were combined into a single variable to avoid losing information of the intensity dimension. In contrast, intensity in Chapter 3 was disregarded for the selection of stimuli and analysis of the data. Future research may expand into investigating the role of perceived intensity in early lifespan development. It may be that there is more complex relationship between emotion and intensity that was masked by combining the separate concepts into one variable.

5.3 Limitations and future research directions

5.3.1 Defining age groups

Within the thesis, the age group of children focused on middle and late childhood between the ages of 5 and 10 year, whereas 11- to 19-year-olds were defined as adolescents (Santrock, 2020; WHO, 2022a). However one limitation could have been the variability of young adult listeners across Chapters 2 and 3. The lower boundary was set to 20 years of age in line with Santrock's (2020) recommendation, whereas the upper age boundary was chosen to match the ages ranges of listeners to those of the speakers. Therefore, in Chapter 2 young adults were participants between 20 and 29 years of age, whereas in Chapter 3, this age range included 20 to 39 year-olds. Whilst this was done to align listener age groups more closely with the age group of the speakers to avoid own-age biases (Amorim et al., 2021), it could have masked effects that should have been detected or inflated others.

For example, Amorim et al. (2021) showed that children had a positive own-age bias, i.e. higher recognition rates detecting affect vocalisations when they were produced by speakers of similar age. This may explain some of the lower recognition rates for children due to the young adult vocalisations used in Chapter 3. However, Amorim et al. (2021) found no own-age biases for adolescents and young adults which is in agreement with others also not reporting own-age biases (Dupuis & Pichora-Fuller, 2015; Morningstar, Ly, et al., 2018; Schirmer et al., 2019). Whilst we cannot entirely rule out own-age biases for our studies due to including only young adult speakers, the chosen age ranges should not have impacted on performance for adolescents and young adults. Given there are only a few studies to date investigating own-age biases in vocal emotion perception, future studies could shed light on the ambiguity by including stimuli produced by children and adolescents. Furthermore, to avoid the pitfalls of allocating age groups (see Chapter 1 for a more detailed explanation), we did analyse the data predominantly with mixed-effect models treating listener age as a continuous variable. This allowed us to observe fine-tuned developmental trajectories independent of allocating listeners into pre-defined age groups.

5.3.2 Validation attempts

The validation methods varied considerably between the three experimental chapters. For personality research, one option is to select stimuli based on prior ratings (e.g. Caulfield et al., 2016; Ewing et al., 2015). The other option is to get listeners to rate the stimuli as they see fit, and interrater reliability scores will subsequently determine how well listeners agreed with one another (e.g. Mahrholz et al., 2018; McAleer et al., 2014). The outcomes/ ratings of the initial study may subsequently serve as a baseline for future studies. In Chapter 2, the selection of stimuli was based on prior perception ratings obtained by McAleer et al. (2014).

For emotion research, however, there are two distinct methodological approaches of validating vocal affect portrayals. The first option is to ask the listener which emotion they perceive, and select a choice from a number of emotion categories provided. An accuracy rating is subsequently computed by comparing these ratings to the speaker's intended emotion category (e.g.

Lassalle et al., 2019). The second option is to provide the listener with additional information of the actor's intended emotion category and ask whether the emotion portrayal is a) recognisable and b) authentic (i.e. genuine). A computed composite score and/or cut-off thresholds of recognisability and authenticity will subsequently inform which stimuli to include in future studies (e.g. Banse & Scherer, 1996; Morningstar, Ly, et al., 2018).

In Chapter 3, three lab members opted to validate the 312 stimuli via the recognisability and authenticity approach. Despite choosing the stimuli with the highest composite score, some of the selected stimuli were not very well recognised by a large group of participants within the Glasgow Science Centre. Having only three lab members available for pre-validation is certainly a limitation compared to larger sample-sized and higher powered validation attempts. However, this method was agreed upon due to time-restrictions, and favoured over the lead researcher selecting stimuli without any further input. Still, it is not uncommon for researchers in the field to pre-validate stimuli with a small number of raters (e.g. Amorim et al., 2021; Morningstar, Ly, et al., 2018).

Given the low recognition rates for some of the stimuli, we may want to argue that these two validation approaches are potentially evaluating different concepts. In both methods, the listener's emotional response is getting compared to the intended emotion category of the actor. Yet, the first approach measures the listener's perception in accordance with the canonical emotions happiness, sadness, anger, surprise, fear, and disgust (Ekman & Friesen, 1971), whereas the second method requires the listener to evaluate whether they agree with the emotion represented. The remaining question is therefore how these two validation approaches compare, and which method should be favoured to select reliable stimuli in future research.

In order to address the caveats outlined in a future study, we collected listener's judgements via both the emotion perception method, as well as via recognisability and authenticity ratings for the Glasgow vocal emotion and personality corpus (Chapter 4). Additionally, in a separate study, we gathered free response ratings from 100 participants (again these ratings were omitted from the thesis due to ongoing analysis) to record listener's "true" perception

regardless of prior knowledge about the speaker's intentions or having a limited number of response categories available (e.g. Elfenbein et al., 2021; Nelson & Russell, 2011). One proposed analysis method for future research could be to explore whether intended emotion, perceived emotion, or authenticity, recognisability or composite scores thereof would serve as the best predictor of free-response ratings. The information would provide researchers with sufficient knowledge to select well-recognised stimuli in future studies.

5.3.3 Methodological considerations

A further limitation in this thesis may have been the estimating of sample sizes which is difficult for mixed effects models. There are tools available when predictors are categorical and designs are fairly simple (e.g. Westfall, 2016). However, challenges remain for continuous predictors and/or more complex designs, especially when a random-effects structure for participants and stimuli is included in the model. Recently, power analysis via simulation methods has been suggested as a way forward (DeBruine, 2021; DeBruine & Barr, 2021). However, those power tools are more successful when based on reliable information/priors of population estimates.

When studying the developmental trajectories of trustworthiness (Chapter 2), there were no reliable priors available. Previous research had studied how trustworthiness develops across the early lifespan, however with designs and stimuli quite different from the ones employed here (e.g. face stimuli, economic games, aggregated scores without random effect structure, etc.). Here, we employed data simulation to conduct sensitivity power analysis as recommended by DeBruine and colleagues (e.g. DeBruine, 2021; DeBruine & Barr, 2021) to address the setbacks. Results suggested that the small age effect determined was indeed significantly powered. Contrastingly, the main effect of presentation (i.e. first or second rating) was underpowered and therefore interpreted as due to chance.

Similarly, when conducting power analysis for Chapter 3, previous literature had not analysed early developmental trajectories of vocal emotion recognition paradigms with mixed-effects models including random-effects structures. Due to the absence of population parameters, sample size estimates were based on

categorical age groups and an effect size of 0.45 as suggested by Jake Westfall (Westfall, 2016; Westfall et al., 2014). The question therefore remains whether this study was sufficiently powered. Sensitivity power analysis may have been a way forward once again, however, it was decided against. Due to the complex design and multitude of predictors, each of the models included in Chapter 3 took 24-36 hours computation time. It was therefore not feasible to conduct data simulations with a minimum of 10,000 replications equivalent to Chapter 2. With that being said, this thesis now provides estimates for fixed and random effects that researchers with similar designs can use as priors to determine sample sizes more reliably.

5.3.4 Terminology of gender and sex

In all demographic questionnaires provided, the terminology of "sex" was used. Whilst gender and sex are frequently used interchangeably, there are distinct differences between them. To outline briefly, sex is defined as an expression referring to biological and physiological characteristics, whereas gender relates to socially constructed characteristics (Gender Spectrum, 2019b, February 21; Planned Parenthood, 2022a; Tolland & Evans, 2019; WHO, 2022b). There is great variety between societies as to the particular norms, roles, and behaviours assigned to boys, girls, women, and men, and those social constructs can change over time (Gender Spectrum, 2019b; Planned Parenthood, 2022a; WHO, 2022b). Furthermore, there is gender identity, which is different from gender and sex, and is defined as "a person's deeply felt, internal and individual experience of gender" (WHO, 2022b).

Labels for sex would include female, male, and intersex (Planned Parenthood, 2022a; WHO, 2022b). Gender labels include boy, girl, woman, and man (Gender Spectrum, 2019a; WHO, 2022b) or use labels in relation to masculinity and femininity (Tolland & Evans, 2019, February 21). Gender identity can but does not have to match the sex a person was assigned at birth (Gender Spectrum, 2019b; WHO, 2022b). Therefore, the two broad categories to distinguish between are cis-gender (i.e. people with matching sex and gender identity) and transgender (i.e. people whose sex and gender identity does not match) (HRC, n.d.; Planned Parenthood, 2022b; Tolland & Evans, 2019, February 21). Non-binary is the umbrella term used for people who do not solely identify

themselves as male or female and may include people who identify as both, somewhere in between, or neither (Gender Spectrum, 2019a; HRC, n.d.; Tolland & Evans, 2019, February 21). Many non-binary people identify as transgender, but not all do (HRC, n.d.). Likewise, in a broader sense, non-binary may include labels of agender, bigender, genderqueer, or gender-fluid (Gender Spectrum, 2019a; HRC, n.d.).

As stated above, all demographics questionnaires for speakers and listeners were designed to ask about the sex of the participant but the term might have gotten confused with gender. For the speakers, this did not pose much of a limitation, since all interactions happened face-to-face and any misunderstanding could have been solved on the spot. Regardless, all speakers identified as cis-gender. We are aware, though, that there was room for interpretation of that question for listeners depending on familiarity with the terminology of sex, gender, and gender identity. Some listeners may have answered the question with their social gender or gender identity rather than their biological sex but may have chosen incorrect labels to indicate. For the studies involving online data collection, there was little opportunity to investigate which concept was referred to (unless participants specifically labelled themselves as non-binary). Given the ambiguity of the terminology, the terms sex and gender were used interchangeably within this thesis. Future research should define the concepts of sex, gender, or gender identity more rigorously and investigate specifically which concept contributes to differences in findings.

5.3.5 Expanding the database

The Glasgow vocal emotion and personality corpus was developed in Chapter 4, starting with the socially-relevant word “hello”. However, disgust was one of the emotion categories that did not work very well with the chosen stimulus type. This is not surprising since disgust usually falls short of being recognised in speech scenarios compared to non-verbal emotion recognition (Banse & Scherer, 1996; Belin et al., 2008; Hawk et al., 2009; Laukka et al., 2013; Lausen & Hammerschmidt, 2020; Lausen & Schacht, 2018; Lima et al., 2013; Sauter, Eisner, Calder, & Scott, 2010). In future, we are aiming to expand the Glasgow vocal emotion and personality corpus to incorporate other stimuli types from the same speakers, such as enacted non-verbal vocalisations (similar to the MAV),

vowels, pseudo-sentences, free speech and creative speech scenarios. This would allow us (or others) to investigate the recognition rates for disgust across different stimulus types encoded by the same speakers. Including further stimuli would also enhance the database to make it more attractive to other researchers in the field, but also would enable to us to draw conclusions on how emotion and personality link beyond socially-relevant word stimuli.

5.4 Conclusion

This thesis has made a novel contribution in the field of vocal trait perceptions by studying the developmental trajectories of perceived vocal trustworthiness across the early lifespan. It has further added to the literature of the early developmental course of vocal emotion perception when employing socially-relevant stimuli. In addition, this thesis has produced the Glasgow vocal emotion and personality corpus, a validated database on a variety of personality and affect measures. Making this database open-access, not only promotes open-science practices, but provides researchers with the opportunity to study emotion or personality with ecologically-valid stimuli. Furthermore, this corpus builds the foundation of studying emotion and personality in unison in the future to understand the complexity of social perceptions from voices better.

References

Aguert, M., Laval, V., Lacroix, A., Gil, S., & Le Bigot, L. (2013). Inferring emotions from speech prosody: Not so easy at age five. *PLoS ONE, 8*(12), Article e83657. https://doi.org/10.1371/journal.pone.0083657

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. https://doi.org/10.1111/bjdp.12097

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *Bmj, 332*(7549), 1080. https://doi.org/10.1136/bmj.332.7549.1080

Amir, O., & Levine-Yundof, R. (2013). Listeners' Attitude Toward People With Dysphonia. *Journal of Voice, 27*(4), Article 524.e1. https://doi.org/10.1016/j.jvoice.2013.01.015

Amorim, M., Anikin, A., Mendes, A. J., Lima, C. F., Kotz, S. A., & Pinheiro, A. P. (2021). Changes in vocal emotion recognition across the life span. *Emotion, 21*(2), 315-325. https://doi.org/10.1037/emo0000692

Apicella, C. L., & Feinberg, D. R. (2009). Voice pitch alters mate-choice-relevant perception in hunter-gatherers. *Proceedings of the Royal Society B-Biological Sciences, 276*(1659), 1077-1082. https://doi.org/10.1098/rspb.2008.1542

Arain, M., Haque, M., Johal, L., Mathur, P., Nel, W., Rais, A., Sandhu, R., & Sharma, S. (2013). Maturation of the adolescent brain. *Neuropsychiatric Disease and Treatment, 9*, 449-461. https://doi.org/10.2147/NDT.S39776

Archer, J. (2004). Sex Differences in Aggression in Real-World Settings: A Meta-Analytic Review. *Review of General Psychology, 8*(4), 291-322. https://doi.org/10.1037/1089-2680.8.4.291

Arnett, J. J. (2016a). Does Emerging Adulthood Theory Apply Across Social Classes? National Data on a Persistent Question. *Emerging Adulthood, 4*(4), 227-235. https://doi.org/10.1177/2167696815613000

Arnett, J. J. (2016b). Emerging Adulthood and Social Class: Rejoinder to Furstenberg, Silva, and du Bois-Reymond. *Emerging Adulthood, 4*(4), 244-247. https://doi.org/10.1177/2167696815627248

Aronovitch, C. D. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *The Journal of Social Psychology, 99*(2), 207-220. https://doi.org/10.1080/00224545.1976.9924774

Audacity Team. (2021). *Audacity®: Free Audio Editor and Recorder*. (Version 2.3.0) Audacity® software is copyright © 1999-2021 Audacity Team. The name Audacity® is a registered trademark. https://audacityteam.org/

Ayala, F. J. (2009). Darwin and the scientific method. *Proceedings of the National Academy of Sciences, 106*(Supplement 1), 10033–10039. https://doi.org/10.1073/pnas.0901404106

Babel, M., McGuire, G., & King, J. (2014). Towards a More Nuanced View of Vocal Attractiveness. *PLoS ONE, 9*(2), e88616. https://doi.org/10.1371/journal.pone.0088616

Bailey, P. E., Szczap, P., McLennan, S. N., Slessor, G., Ruffman, T., & Rendell, P. G. (2015). Age-related similarities and differences in first impressions of trustworthiness. *Cognition & Emotion*, 1-10. https://doi.org/10.1080/02699931.2015.1039493

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614-636. https://doi.org/10.1037/0022-3514.70.3.614

Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion, 9*(5), 691-704. https://doi.org/10.1037/a0017088

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion, 12*(5), 1161. https://doi.org/10.1037/a0025827

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion, 6*(2), 269-278. https://doi.org/10.1037/1528-3542.6.2.269

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Baugh, J. (2000). Racial identifications by speech. *American Speech, 75*(4), 362-364. https://doi.org/10.1215/00031283-75-4-362

Baus, C., McAleer, P., Marcoux, K., Belin, P., & Costa, A. (2019). Forming social impressions from voices in native and foreign languages. *Scientific Reports, 9*, Article 414 (2019). https://doi.org/10.1038/s41598-018-36518-6

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior research methods, 40*(2), 531-539. https://doi.org/10.3758/BRM.40.2.531

Berk, L. E. (2017). *Development through the lifespan* (7th ed.). Pearson.

Berry, D. S. (1990). Vocal attractiveness and vocal babyishness: Effects on stranger, self, and friend impressions. *Journal of Nonverbal Behavior, 14*(3), 141-153. https://doi.org/10.1007/bf00996223

Blakemore, S.-J. (2018). *Inventing ourselves : the secret life of the teenage brain*. Doubleday.

Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer*. (Version 6.1.55) http://www.praat.org/

Bolker, B., & Robinson, D. (2020). *broom.mixed: Tidying Methods for Mixed Models*. (Version 0.2.7) https://CRAN.R-project.org/package=broom.mixed

Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour, 82*(1), 55-59. https://doi.org/10.1016/j.anbehav.2011.03.024

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *Peerj, 8*, e9414. https://doi.org/10.7717/peerj.9414

Brosgole, L., & Weisman, J. (1995). Mood recognition across the ages. *International Journal of Neuroscience, 82*, 169-189. https://doi.org/10.3109/00207459508999800

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., Kawahara, H., & Belin, P. (2010). Vocal attractiveness increases by averaging. *Current Biology, 20*(2), 116-120. https://doi.org/10.1016/j.cub.2009.11.034

Brückl, M., & Heuer, F. (2021). *irrNA: Coefficients of interrater reliability: Generalized for randomly incomplete datasets*. (Version 0.2.2) https://cran.r-project.org/web/packages/irrNA/

Bryant, G., & Barrett, H. C. (2008). Vocal Emotion Recognition Across Disparate Cultures. *Journal of Cognition and Culture, 8*(1-2), 135-148. https://doi.org/10.1163/156770908X289242

Burgoon, J. K., Guerrero, L. K., & Floyd, K. (2010). *Nonverbal communication*. Routledge.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech [Conference Paper]. *Proceedings Interspeech 2005, 5*, 1517-1520. https://doi.org/10.21437/Interspeech.2005-446

Buss, D. M. (1989). Conflict between the sexes: Strategic interference and the evocation of anger and upset. *Journal of Personality and Social Psychology, 56*(5), 735-747. https://doi.org/10.1037/0022-3514.56.5.735

Carstensen, L. L. (2006). The influence of a sense of time on human development. *Science, 312*(5782), 1913-1915. https://doi.org/10.1126/science.1127488

Carstensen, L. L., & DeLiema, M. (2018). The positivity effect: a negativity bias in youth fades with age. *Current Opinion in Behavioral Sciences, 19*, 7-12. https://doi.org/10.1016/j.cobeha.2017.07.009

Carstensen, L. L., Isaacowitz, D. M., & Charles, S. T. (1999). Taking time seriously. A theory of socioemotional selectivity. *American Psychologist, 54*(3), 165-181. https://doi.org/10.1037//0003-066x.54.3.165

Castle, E., Eisenberger, N. I., Seeman, T. E., Moons, W. G., Boggero, I. A., Grinblatt, M. S., & Taylor, S. E. (2012). Neural and behavioral bases of age differences in perceptions of trust. *Proceedings of the National Academy of Sciences, 109*(51), 20848-20852. https://doi.org/10.1073/pnas.1218518109

Castro, S. L., & Lima, C. F. (2010). Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody. *Behavior research methods, 42*(1), 74-81. https://doi.org/10.3758/BRM.42.1.74

Caulfield, F., Ewing, L., Bank, S., & Rhodes, G. (2016). Judging trustworthiness from faces: Emotion cues modulate trustworthiness judgments in young children. *British Journal of Psychology, 107*(3), 503-518. https://doi.org/10.1111/bjop.12156

CDC. (2021, February 22). *Positive Parenting Tips*. Retrieved February 5, 2022, from https://www.cdc.gov/ncbddd/childdevelopment/positiveparenting/index.html

Cepon-Robins, T. J., Blackwell, A. D., Gildner, T. E., Liebert, M. A., Urlacher, S. S., Madimenos, F. C., Eick, G. N., Snodgrass, J. J., & Sugiyama, L. S. (2021). Pathogen disgust sensitivity protects against infection in a high pathogen environment. *Proceedings of the National Academy of Sciences, 118*(8), Article e2018552118. https://doi.org/10.1073/pnas.2018552118

Champely, S. (2020). *pwr: Basic functions for power analysis.* (Version 1.3-0) https://CRAN.R-project.org/package=pwr

Chaplin, T. M. (2014). Gender and Emotion Expression: A Developmental Contextual Perspective. *Emotion Review, 7*(1), 14-21. https://doi.org/10.1177/1754073914544408

Charlesworth, T. E. S., Hudson, S.-k. T. J., Cogsdill, E. J., Spelke, E. S., & Banaji, M. R. (2019). Children use targets' facial appearance to guide and predict social behavior. *Developmental Psychology, 55*(7), 1400-1413. https://doi.org/10.1037/dev0000734

Chen, C., & Jack, R. E. (2017). Discovering cultural differences (and similarities) in facial expressions of emotion. *Emotion, 17*, 61-66. https://doi.org/10.1016/j.copsyc.2017.06.010

Chen, X., Yang, J., Gan, S., & Yang, Y. (2012). The Contribution of Sound Intensity in Vocal Emotion Perception: Behavioral and Electrophysiological Evidence. *PLoS ONE, 7*(1), Article e30278. https://doi.org/10.1371/journal.pone.0030278

Christensen, R. H. B. (2015). *ordinal: Regression Models for Ordinal Data*. (Version 2019.12-10) https://cran.r-project.org/package=ordinal

Chronaki, G., Hadwin, J. A., Garner, M., Maurage, P., & Sonuga-Barke, E. J. S. (2015). The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. *British Journal of Developmental Psychology, 33*(2), 218-236. https://doi.org/10.1111/bjdp.12075

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), Article 8659 (2018). https://doi.org/10.1038/s41598-018-26889-1

Cogsdill, E. J., & Banaji, M. R. (2015). Face-trait inferences show robust child–adult agreement: Evidence from three types of faces. *Journal of Experimental Social Psychology, 60*, 150-156. https://doi.org/10.1016/j.jesp.2015.05.007

Cogsdill, E. J., Todorov, A. T., Spelke, E. S., & Banaji, M. R. (2014). Inferring Character From Faces: A Developmental Study. *Psychological Science, 25*(5), 1132-1139. https://doi.org/10.1177/0956797614523297

Collignon, O., Girard, S., Gosselin, F., Saint-Amour, D., Lepore, F., & Lassonde, M. (2010). Women process multisensory emotion expressions more efficiently than men. *Neuropsychologia, 48*(1), 220-225. https://doi.org/10.1016/j.neuropsychologia.2009.09.007

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour, 65*(5), 997-1004. https://doi.org/10.1006/anbe.2003.2123

Corr, P. J., & Krupić, D. (2017). Chapter Two - Motivating Personality: Approach, Avoidance, and Their Conflict. In A. J. Elliot (Ed.), *Advances in Motivation Science* (Vol. 4, pp. 39-90). Elsevier. https://doi.org/10.1016/bs.adms.2017.02.003

Cortes, D. S., Tornberg, C., Bänziger, T., Elfenbein, H. A., Fischer, H., & Laukka, P. (2021). Effects of aging on emotion recognition from dynamic multimodal expressions and vocalizations. *Scientific Reports, 11*, Article 2647 (2021). https://doi.org/10.1038/s41598-021-82135-1

Darwin, C. (1872). *The expression of the emotions in man and animals*. John Murray. https://pure.mpg.de/rest/items/item_2309885/component/file_2309884/content

DeBruine, L. M. (2021). *faux: Simulation for Factorial Designs*. (Version 1.1.0) https://CRAN.R-project.org/package=faux

DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science, 4*(1), 1-15. https://doi.org/10.1177/2515245920965119

DeBruine, L. M., Lai, R., Jones, B. C., Abdullah, R., & Mahrholz, G. (2020). *Experimentum*. (Version v.0.2) Zenodo. https://doi.org/10.5281/zenodo.2634355

Deffenbacher, J. L., Oetting, E. R., Lynch, R. S., & Morris, C. D. (1996). The expression of anger and its consequences. *Behaviour Research and Therapy, 34*(7), 575-590. https://doi.org/10.1016/0005-7967(96)00018-6

Demenescu, L. R., Mathiak, K. A., & Mathiak, K. (2014). Age- and Gender-Related Variations of Emotion Recognition in Pseudowords and Faces. *Experimental Aging Research, 40*(2), 187-207. https://doi.org/10.1080/0361073x.2014.882210

Doherty, C. P., Fitzsimons, M., Asenbauer, B., & Staunton, H. (1999). Discrimination of prosody and music by normal children. *European Journal of Neurology, 6*(2), 221-226. https://doi.org/10.1111/j.1468-1331.1999.tb00016.x

Dosenbach, N. U. F., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., Nelson, S. M., Wig, G. S., Vogel, A. C., Lessov-Schlaggar, C. N., Barnes, K. A., Dubis, J. W., Feczko, E., Coalson, R. S., Pruett, J. R., Barch, D. M., Petersen, S. E., & Schlaggar, B. L. (2010). Prediction of Individual Brain Maturity Using fMRI. *Science, 329*(5997), 1358-1361. https://doi.org/10.1126/science.1194144

du Bois-Reymond, M. (2016). Emerging Adulthood Theory Under Scrutiny. *Emerging Adulthood, 4*(4), 242-243. https://doi.org/10.1177/2167696815614422

Dupuis, K., & Pichora-Fuller, M. K. (2015). Aging Affects Identification of Vocal Emotions in Semantically Neutral Sentences. *Journal of speech, language, and hearing research, 58*(3), 1061-1076. https://doi.org/10.1044/2015_JSLHR-H-14-0256

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*(2), 124-129. https://doi.org/10.1037/h0030377

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*, 1-18. https://doi.org/10.1177/01461672211029786

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience, 19*(9), 1508-1519. <Go to WoS>://WOS:000257153000004

Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of Emotional Information in Voice-Sensitive Cortices. *Current Biology, 19*(12), 1028-1033. https://doi.org/10.1016/j.cub.2009.04.054

Ewing, L., Caulfield, F., Read, A., & Rhodes, G. (2015). Perceived trustworthiness of faces drives trust behaviour in children. *Developmental Science, 18*(2), 327-334. https://doi.org/10.1111/desc.12218

Ewing, L., Sutherland, C. A. M., & Willis, M. L. (2019). Children show adult-like facial appearance biases when trusting others. *Developmental Psychology, 55*(8), 1694–1701. https://doi.org/10.1037/dev0000747

Fairbanks, G. (1960). The rainbow passage. In *Voice and articulation drillbook* (2 ed., pp. 124-139). Harper & Row.

Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2008). The Role of Femininity and Averageness of Voice Pitch in Aesthetic Judgments of Women's Voices. *Perception, 37*(4), 615-623. https://doi.org/10.1068/p5514

Ferdenzi, C., Delplanque, S., Mehu-Blantar, I., Da Paz Cabral, K. M., Domingos Felicio, M., & Sander, D. (2015). The Geneva Faces and Voices (GEFAV) database. *Behavior research methods, 47*(4), 1110-1121. https://doi.org/10.3758/s13428-014-0545-0

Ferdenzi, C., Patel, S., Mehu-Blantar, I., Khidasheli, M., Sander, D., & Delplanque, S. (2013). Voice attractiveness: Influence of stimulus duration and type. *Behavior research methods, 45*(2), 405-413. https://doi.org/10.3758/s13428-012-0275-0

Fessler, D. M. T., Pillsworth, E. G., & Flamson, T. J. (2004). Angry men and disgusted women: An evolutionary approach to the influence of emotions on risk taking. *Organizational Behavior and Human Decision Processes, 95*(1), 107-123. https://doi.org/10.1016/j.obhdp.2004.06.006

Fischer, A. H., & Evers, C. (2010). Anger in the Context of Gender. In M. Potegal, G. Stemmler, & C. Spielberger (Eds.), *International Handbook of Anger: Constituent and Concomitant Biological, Psychological, and Social Processes* (pp. 349-360). Springer New York. https://doi.org/10.1007/978-0-387-89676-2_20

Fischer, A. H., & Rodriguez Mosquera, P. M. (2001). What concerns men? Women or other men?: A critical appraisal of the evolutionary theory of gender differences. *Psychology, Evolution & Gender, 3*(1), 5-26. https://doi.org/10.1080/14616660110049564

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77-83. https://doi.org/10.1016/j.tics.2006.11.005

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), 238-252. https://doi.org/10.1037/0012-1649.43.1.238

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third ed.). Sage Publications. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fox, J., Weisberg, S., & Price, B. (2021). *car: Companion to Applied Regression.* (Version 3.0-11) https://cran.r-project.org/package=car

Friend, M. (2000). Developmental changes in sensitivity to vocal paralanguage. *Developmental Science, 3*(2), 148-162. https://doi.org/10.1111/1467-7687.00108

Furstenberg, F. (2016). Social Class and Development in Early Adulthood: Some Unsettled Issues. *Emerging Adulthood, 4*(4), 236-238. https://doi.org/10.1177/2167696815625142

Gender Spectrum. (2019a). *The Language of Gender*. Retrieved February 2015, 2022, from https://www.genderspectrum.org/articles/language-of-gender

Gender Spectrum. (2019b). *Understanding Gender*. Retrieved February 15, 2022, from https://www.genderspectrum.org/articles/understanding-gender

Giles, H., & Billings, A. C. (2004). Assessing Language Attitudes: Speaker Evaluation Studies. In A. Davies & C. Elder (Eds.), *The Handbook of Applied Linguistics* (pp. 187-209). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470757000.ch7

Goddard, C. (2014). On "Disgust". In F. Baider & G. Cislaru (Eds.), *Linguistic approaches to emotions in context* (Vol. 241, pp. 73-97). John Benjamins Publishing Company.

Gold, R., Butler, P., Revheim, N., Leitman, D. I., Hansen, J. A., Gur, R. C., Kantrowitz, J. T., Laukka, P., Juslin, P. N., Silipo, G. S., & Javitt, D. C. (2012). Auditory emotion recognition impairments in schizophrenia: relationship to acoustic features and cognition. *The American Journal Of Psychiatry, 169*(4), 424-432. https://doi.org/10.1176/appi.ajp.2011.11081230

Griffiths, S., Penton-Voak, I. S., Jarrold, C., & Munafò, M. R. (2015). No Own-Age Advantage in Children's Recognition of Emotion on Prototypical Faces of Different Ages. *PLoS ONE, 10*(5), Article e0125256. https://doi.org/10.1371/journal.pone.0125256

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software, 40*(3), 1-25. https://doi.org/10.18637/jss.v040.i03

Grosbras, M.-H., Ross, P. D., & Belin, P. (2018). Categorical emotion recognition from voice improves during childhood and adolescence. *Scientific Reports, 8*, Article 14791. https://doi.org/10.1038/s41598-018-32868-3

Gudicha, D. W., Schmittmann, V. D., & Vermunt, J. K. (2017). Statistical power of likelihood ratio and Wald tests in latent class models with covariates. *Behavior research methods, 49*(5), 1824-1837. https://doi.org/10.3758/s13428-016-0825-y

Hall, J. A., Gunnery, S. D., Letzring, T. D., Carney, D. R., & Colvin, C. R. (2017). Accuracy of Judging Affect and Accuracy of Judging Personality: How and When Are They Related? *Journal of Personality, 85*(5), 583-592. https://doi.org/10.1111/jopy.12262

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Harrell Jr, F. E. (2021). *Hmisc: Harrell Miscellaneous*. (Version 4.5-0) https://CRAN.R-project.org/package=Hmisc

Hawk, S. T., Van Kleef, G. A., Fischer, A. H., & Van Der Schalk, J. (2009). "Worth a thousand words": absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion, 9*(3), 293-305. https://doi.org/10.1037/a0015178

Hawkley, L. C., & Capitanio, J. P. (2015). Perceived social isolation, evolutionary fitness and health outcomes: a lifespan approach. *Philosophical Transactions Of The Royal Society Of London. Series B, Biological Sciences, 370*(1669), 17-21. https://doi.org/10.1098/rstb.2014.0114

Herring, S. C., & Dainas, A. R. (2020). Gender and Age Influences on Interpretation of Emoji Functions. *ACM Transactions on Social Computing, 3*(2), Article 10. https://doi.org/10.1145/3375629

Hills, P. J., & Willis, S. F. L. (2016). Children view own-age faces qualitatively differently to other-age faces. *Journal of Cognitive Psychology, 28*(5), 601-610. https://doi.org/10.1080/20445911.2016.1164710

Holz, N., Larrouy-Maestri, P., & Poeppel, D. (2021). The paradoxical role of emotional intensity in the perception of vocal affect. *Scientific Reports, 11*(1), Article 9663. https://doi.org/10.1038/s41598-021-88431-0

HRC. (n.d.). *Human Rights Campain - Glossary*. Retrieved February 15, 2022, from https://www.hrc.org/resources/glossary-of-terms

Hughes, S. M., & Rhodes, B. C. (2010). Making age assessments based on voice: The impact of the reproductive viability of the speaker. *Journal of Social, Evolutionary, and Cultural Psychology, 4*(4), 290-304. https://doi.org/10.1037/h0099282

Hunter, E. M., Phillips, L. H., & MacPherson, S. E. (2010). Effects of age on cross-modal emotion perception. *Psychology and Aging, 25*(4), 779-787. https://doi.org/10.1037/a0020528

IEEE Recommended Practice for Speech Quality Measurements. (1969). *IEEE Transactions on Audio and Electroacoustics, 17*(3), 225-246. https://doi.org/10.1109/TAU.1969.1162058

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences, 109*(19), 7241-7244. https://doi.org/10.1073/pnas.1200155109

Jack, R. E., Sun, W., Delis, I., Garrod, O. G. B., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General, 145*(6), 708-730. https://doi.org/10.1037/xge0000162

Jessen, S., & Grossmann, T. (2016). Neural and Behavioral Evidence for Infants' Sensitivity to the Trustworthiness of Faces. *Journal of Cognitive Neuroscience, 28*(11), 1728-1736. https://doi.org/10.1162/jocn_a_00999

Jiang, X., & Pell, M. D. (2015). On how the brain decodes vocal cues about speaker confidence. *Cortex, 66*, 9-34. https://doi.org/10.1016/j.cortex.2015.02.002

Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., & Vukovic, J. (2008). Integrating cues of social interest and voice pitch in men's preferences for women's voices. *Biology Letters, 4*(2), 192-194. https://doi.org/10.1098/rsbl.2007.0626

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion, 1*(4), 381-412. https://doi.org/10.1037/1528-3542.1.4.381

Juslin, P. N., Laukka, P., & Bänziger, T. (2018). The Mirror to Our Soul? Comparisons of Spontaneous and Posed Vocal Expression of Emotion. *Journal of Nonverbal Behavior, 42*, 1-40. https://doi.org/10.1007/s10919-017-0268-x

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. https://doi.org/10.3758/s13423-019-01701-x

Kassambara, A. (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots.* (Version 0.4.0) https://CRAN.R-project.org/package=ggpubr

Kassambara, A. (2021). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. (Version 0.7.0) https://CRAN.R-project.org/package=rstatix

Kawahara, M., Sauter, D. A., & Tanaka, A. (2017). Impact of Culture on the Development of Multisensory Emotion Perception. *Proc. The 14th International Conference on Auditory-Visual Speech Processing*, 109-114. https://doi.org/10.21437/AVSP.2017-21

Kawahara, M., Sauter, D. A., & Tanaka, A. (2021). Culture shapes emotion perception from faces and voices: changes over development. *Cognition and Emotion*, 1-12. https://doi.org/10.1080/02699931.2021.1922361

Keshtiari, N., & Kuhlmann, M. (2016). The Effects of Culture and Gender on the Recognition of Emotional Speech: Evidence from Persian Speakers Living in a Collectivist Society. *International Journal of Society, Culture & Language, 4*(2), 71-86. http://www.ijscl.net/article_19785.html

Kilford, E. J., Garrett, E., & Blakemore, S.-J. (2016). The development of social cognition in adolescence: An integrated perspective. *Neuroscience & Biobehavioral Reviews, 70*, 106-120. https://doi.org/10.1016/j.neubiorev.2016.08.016

Kinghorn, A., Shanaube, K., Toska, E., Cluver, L., & Bekker, L.-G. (2018). Defining adolescence: priorities from a global health perspective. *The Lancet Child & Adolescent Health, 2*(5), Article e10. https://doi.org/10.1016/S2352-4642(18)30096-8

Kitzing, P. (1986). LTAS criteria pertinent to the measurement of voice quality. *Journal of Phonetics, 14*(3), 477-482. https://doi.org/10.1016/S0095-4470(19)30693-X

Klofstad, C. A., & Anderson, R. C. (2018). Voice pitch predicts electability, but does not signal leadership ability. *Evolution and Human Behavior, 39*(3), 349-354. https://doi.org/10.1016/j.evolhumbehav.2018.02.007

Klofstad, C. A., Anderson, R. C., & Nowicki, S. (2015). Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. *PLoS ONE, 10*(8), Article e0133779. https://doi.org/10.1371/journal.pone.0133779

Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B-Biological Sciences, 279*(1738), 2698-2704. https://doi.org/10.1098/rspb.2012.0311

Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., & Okubo, Y. (2013). Cross-Cultural Differences in the Processing of Non-Verbal Affective Vocalizations by Japanese and Canadian Listeners. *Frontiers in Psychology, 4*, 105. https://doi.org/10.3389/fpsyg.2013.00105

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*(6), 618-625. https://doi.org/10.1016/S0022-1031(02)00510-3

Kreiman, J., & Sidtis, D. (2011). Linguistic Uses of Voice Quality: How Voice Signals Linguistic and Pragmatic Aspects of Communication. In J. Kreiman & D. Sidtis (Eds.), *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (pp. 260-301). Wiley Online Books. https://doi.org/10.1002/9781444395068.ch8

Kushins, E. R. (2014). Sounding Like Your Race in the Employment Process: An Experiment on Speaker Voice, Race Identification, and Stereotyping. *Race and Social Problems, 6*(3), 237-248. https://doi.org/10.1007/s12552-014-9123-4

Lally, M., & Valentine-French, S. (2019). *Lifespan Development: A Psychological Perspective* (2nd ed.). Open Textbook Library (CC BY-NC-SA 3.0). https://open.umn.edu/opentextbooks/textbooks/540

Lambrecht, L., Kreifelts, B., & Wildgruber, D. (2012). Age-related decrease in recognition of emotional facial and prosodic expressions. *Emotion, 12*(3), 529-539. https://doi.org/10.1037/a0026827

Lambrecht, L., Kreifelts, B., & Wildgruber, D. (2014). Gender differences in emotion recognition: Impact of sensory modality and emotional category. *Cognition & Emotion, 28*(3), 452-469. https://doi.org/10.1080/02699931.2013.837378

Lassalle, A., Pigat, D., O'Reilly, H., Berggen, S., Fridenson-Hayo, S., Tal, S., Elfström, S., Råde, A., Golan, O., Bölte, S., Baron-Cohen, S., & Lundqvist, D. (2019). The EU-Emotion Voice Database. *Behavior research methods, 51*(2), 493-506. https://doi.org/10.3758/s13428-018-1048-1

Latinus, M., & Belin, P. (2011a). Anti-Voice Adaptation Suggests Prototype-Based Coding of Voice Identity. *Frontiers in Psychology, 2*, Article 175. https://doi.org/10.3389/fpsyg.2011.00175

Latinus, M., & Belin, P. (2011b). Human voice perception. *Current Biology, 21*(4), R143-R145. https://doi.org/10.1016/j.cub.2010.12.033

Latinus, M., & Belin, P. (2012). Perceptual Auditory Aftereffects on Voice Identity Using Brief Vowel Stimuli. *PLoS ONE, 7*(7), Article e41384. https://doi.org/10.1371/journal.pone.0041384

Laukka, P., Bänziger, T., Israelsson, A., Cortes, D. S., Tornberg, C., Scherer, K. R., & Fischer, H. (2021). Investigating individual differences in emotion recognition ability using the ERAM test. *Acta Psychologica, 220*, 103422. https://doi.org/10.1016/j.actpsy.2021.103422

Laukka, P., & Elfenbein, H. A. (2021). Cross-Cultural Emotion Recognition and In-Group Advantage in Vocal Expression: A Meta-Analysis. *Emotion Review, 13*(1), 3-11. https://doi.org/10.1177/1754073919897295

Laukka, P., Elfenbein, H. A., Chui, W., Thingujam, N. S., Iraki, F. K., Rockstuhl, T., & Althoff, J. (2010). Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. In L. Devillers, B. Schuller, R. Cowie, E. Douglas-Cowie, & A. Batliner (Eds.), *Proceedings of the LREC 2010 Workshop on Corpora for Research on Emotion and Affect* (pp. 53-57). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2010/index.html

Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Iraki, F. K. e., Rockstuhl, T., & Thingujam, N. (2013). Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in Psychology, 4*, Article 353. https://doi.org/10.3389/fpsyg.2013.00353

Laukka, P., & Juslin, P. N. (2007). Similar patterns of age-related differences in emotion recognition from speech and music. *Motivation and Emotion, 31*(3), 182-191. https://doi.org/10.1007/s11031-007-9063-z

Lausen, A., & Hammerschmidt, K. (2020). Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters. *Humanities and Social Sciences Communications, 7*(2), 1-17. https://doi.org/10.1057/s41599-020-0499-z

Lausen, A., & Schacht, A. (2018). Gender Differences in the Recognition of Vocal Emotions. *Frontiers in Psychology, 9*, Article 882. https://doi.org/10.3389/fpsyg.2018.00882

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review, 26*, 90-102. https://doi.org/10.3758/s13423-018-1497-7

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications, 10*, Article 2404 (2019). https://doi.org/10.1038/s41467-019-10295-w

Lavan, N., Mileva, M., Burton, A. M., Young, A. W., & McGettigan, C. (2021). Trait evaluations of faces and voices: Comparing within-and between-person variability. *Journal of Experimental Psychology: General*, Article Advance online publication. https://doi.org/10.1037/xge0001019

Lavan, N., Mileva, M., & McGettigan, C. (2020). How does familiarity with a voice affect trait judgements? *British Journal of Psychology, 112*, 282-300. https://doi.org/10.1111/bjop.12454

Laver, J. (1994). *Principles of Phonetics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139166621

Lenth, R. V. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. (Version 1.6.3) https://CRAN.R-project.org/package=emmeans

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology, 81*(1), 146-159. https://doi.org/10.1037/0022-3514.81.1.146

Lima, C. F., Alves, T., Scott, S. K., & Castro, S. L. (2014). In the Ear of the Beholder: How Age Shapes Emotion Processing in Nonverbal Vocalizations. *Emotion, 14*(1), 145-160. https://doi.org/10.1037/a0034287

Lima, C. F., Anikin, A., Monteiro, A. C., Scott, S. K., & Castro, S. L. (2019). Automaticity in the recognition of nonverbal emotional vocalizations. *Emotion, 19*(2), 219-233. https://doi.org/10.1037/emo0000429

Lima, C. F., & Castro, S. L. (2011). Emotion recognition in music changes across the adult life span. *Cognition and Emotion, 25*(4), 585-598. https://doi.org/10.1080/02699931.2010.502449

Lima, C. F., Castro, S. L., & Scott, S. K. (2013). When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing. *Behavior research methods, 45*(4), 1234-1245. https://doi.org/10.3758/s13428-013-0324-3

Litt, E., Zhao, S., Kraut, R., & Burke, M. (2020). What Are Meaningful Social Interactions in Today's Media Landscape? A Cross-Cultural Survey. *Social Media + Society, 6*(3), 1-17. https://doi.org/10.1177/2056305120942888

Liuni, M., Ponsot, E., Bryant, G. A., & Aucouturier, J. J. (2020). Sound context modulates perceived vocal emotion. *Behavioural Processes, 172*, Article 104042. https://doi.org/10.1016/j.beproc.2020.104042

Lortie, C. L., Deschamps, I., Guitton, M. J., & Tremblay, P. (2018). Age Differences in Voice Evaluation: From Auditory-Perceptual Evaluation to Social Interactions. *Journal of Speech Language and Hearing Research, 61*(2), 227-245. https://doi.org/10.1044/2017_jslhr-s-16-0202

Lüdecke, D., Makowski, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., Wiernik, B. M., & Arel-Bundock, V. (2021). *performance: Assessment of Regression Models Performance*. (Version 0.8.0) https://CRAN.R-project.org/package=performance

Lyon, P. (2011). To Be or Not To Be: Where Is Self-Preservation in Evolutionary Theory? In B. Callcott & K. Sterelny (Eds.), *The major transitions in evolution revisited* (pp. 105-126). MIT Press. https://doi.org/10.7551/mitpress/9780262015240.003.0007

Mackenzie, C., MacDougall, C., Fane, J., & Gibbs, L. (2018). *Using emoji in research with children and young people: Because we can?* (Vol. 2). World Conference on Qualitative Research. https://www.proceedings.wcqr.info/index.php/wcqr2019/article/view/11

Mahrholz, G., Belin, P., & McAleer, P. (2018). Judgements of a speaker's personality are correlated across differing content and stimulus type. *PLoS ONE, 13*(10), Article e0204991. https://doi.org/10.1371/journal.pone.0204991

Mahrholz, G., McAleer, P., & Norrbo, M. (2020). Developing voice perception: An overview of current research and models in affect and identity recognition in children and adults. *Cognitive Psychology Bulletin, - Issue 5 Spring 2020,* 58-65. https://doi.org/10.31234/osf.io/eq2yx

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *44*(2), 314-324. https://doi.org/10.3758/s13428-011-0168-7

MATLAB. (2016). *Version 9.1 (R2016b)*. Natick, Massachusetts: The MathWorks Inc.

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. *PLoS ONE, 9*(3), Article e90779. https://doi.org/10.1371/journal.pone.0090779

McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review, 90*(3), 215-238. https://doi.org/10.1037/0033-295X.90.3.215

Mennella, R., Vilarem, E., & Grèzes, J. (2020). Rapid approach-avoidance responses to emotional displays reflect value-based decisions: Neural evidence from an EEG study. *Neuroimage, 222*, 117253. https://doi.org/10.1016/j.neuroimage.2020.117253

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2021). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* (Version 1.7-8) https://CRAN.R-project.org/package=e1071

Mildner, V., & Koska, T. (2014). Recognition and production of emotions in children with cochlear implants. *Clinical Linguistics & Phonetics, 28*(7-8), 543-554. https://doi.org/10.3109/02699206.2014.927000

Mileva, M., & Lavan, N. (2022). How quickly can we form a trait impression from voices? *PsyArXiv*. https://doi.org/10.31234/osf.io/zd4un

Min, C. S., & Schirmer, A. (2011). Perceiving verbal and vocal emotions in a second language. *Cognition and Emotion, 25*(8), 1376-1392. https://doi.org/10.1080/02699931.2010.544865

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior, 27*(4), 237-254. https://doi.org/10.1023/A:1027332800296

Moons, W. G., Eisenberger, N. I., & Taylor, S. E. (2010). Anger and fear responses to stress have different biological profiles. *24*(2), 215-219. https://doi.org/10.1016/j.bbi.2009.08.009

Morningstar, M., Dirks, M. A., & Huang, S. (2017). Vocal cues underlying youth and adult portrayals of socio-emotional expressions. *Journal of Nonverbal Behavior, 41*(2), 155-183. https://doi.org/10.1007/s10919-017-0250-7

Morningstar, M., Ly, V. Y., Feldman, L., & Dirks, M. A. (2018). Mid-Adolescents' and Adults' Recognition of Vocal Cues of Emotion and Social Intent: Differences by Expression and Speaker Age. *Journal of Nonverbal Behavior, 42*(2), 237-251. https://doi.org/10.1007/s10919-018-0274-7

Morningstar, M., Nelson, E. E., & Dirks, M. A. (2018). Maturation of vocal emotion recognition: Insights from the developmental and neuroimaging literature. *Neuroscience and biobehavioral reviews, 90*, 221-230. https://doi.org/10.1016/j.neubiorev.2018.04.019

Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child Development, 72*(3), 834-843. https://doi.org/10.1111/1467-8624.00318

Morton, J. B., Trehub, S. E., & Zelazo, P. D. (2003). Sources of inflexibility in 6-year-olds' understanding of emotion in speech. *Child Development, 74*(6), 1857-1868. https://doi.org/10.1046/j.1467-8624.2003.00642.x

Moyse, E., Beaufort, A., & Brédart, S. (2014). Evidence for an own-age bias in age estimation from voices in older persons. *European Journal of Ageing, 11*(3), 241-247. https://doi.org/10.1007/s10433-014-0305-0

Müller, K. (2021). *hms: Pretty Time of Day.* (Version 1.1.0) https://cran.r-project.org/package=hms

Nash, J. C. (2014). On Best Practice Optimization Methods in R. *Journal of Statistical Software, 60*(2), 1-14. https://doi.org/10.18637/jss.v060.i02

Nash, J. C., & Varadhan, R. (2011). Unifying Optimization Algorithms to Aid Software System Users: optimx for R. *Journal of Statistical Software, 43*(9), 1-14. https://doi.org/10.18637/jss.v043.i09

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. https://doi.org/10.1016/j.jecp.2011.03.014

O'Reilly, H., Pigat, D., Fridenson, S., Berggren, S., Tal, S., Golan, O., Bölte, S., Baron-Cohen, S., & Lundqvist, D. (2016). The EU-Emotion Stimulus Set: A validation study. *48*(2), 567-576. https://doi.org/10.3758/s13428-015-0601-4

Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., & Sorokowska, A. (2017). Voice-based assessments of trustworthiness, competence, and warmth in

blind and sighted adults. *Psychonomic Bulletin & Review, 24*(3), 856-862. https://doi.org/10.3758/s13423-016-1146-y

ONS. (2016, February 22). *Why are more young people living with their parents?* Retrieved February 5, 2022, from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/articles/whyaremoreyoungpeoplelivingwiththeirparents/2016-02-22

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087-11092. https://doi.org/10.1073/pnas.0805664105

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22-27. https://doi.org/10.1016/j.jbef.2017.12.004

Palmquist, C. M., Cheries, E. W., & DeAngelis, E. R. (2020). Looking smart: Preschoolers' judgements about knowledge based on facial appearance. *British Journal of Developmental Psychology, 38*(1), 31-41. https://doi.org/10.1111/bjdp.12303

Parks, S. L., & Clancy Dollinger, S. (2014). The positivity effect and auditory recognition memory for musical excerpts in young, middle-aged, and older adults. *Psychomusicology: Music, Mind, and Brain, 24*(4), 298-308. https://doi.org/10.1037/pmu0000079

Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology, 87*(1), 93-98. https://doi.org/10.1016/j.biopsycho.2011.02.010

Patton, G. C., Olsson, C. A., Skirbekk, V., Saffery, R., Wlodek, M. E., Azzopardi, P. S., Stonawski, M., Rasmussen, B., Spry, E., Francis, K., Bhutta, Z. A., Kassebaum, N. J., Mokdad, A. H., Murray, C. J. L., Prentice, A. M., Reavley, N., Sheehan, P., Sweeny, K., Viner, R. M., & Sawyer, S. M. (2018). Adolescence and the next generation. *Nature, 554*(7693), 458-466. https://doi.org/10.1038/nature25759

Paulmann, S., & Pell, M. D. (2010). Dynamic emotion processing in Parkinson's disease as a function of channel availability. *Journal Of Clinical And Experimental Neuropsychology, 32*(8), 822-835. https://doi.org/10.1080/13803391003596371

Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). How aging affects the recognition of emotional speech. *Brain and Language, 104*(3), 262-269. https://doi.org/10.1016/j.bandl.2007.03.002

Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition and Emotion, 28*(2), 230-244. https://doi.org/10.1080/02699931.2013.812033

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *51*, 195-203. https://doi.org/10.3758/s13428-018-01193-y

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics, 37*(4), 417-435. https://doi.org/10.1016/j.wocn.2009.07.005

Pell, M. D., & Skorup, V. (2008). Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication, 50*(6), 519-530. https://doi.org/10.1016/j.specom.2008.03.006

Pinheiro, A. P., Anikin, A., Conde, T., Sarzedas, J., Chen, S., Scott, S. K., & Lima, C. F. (2021). Emotional authenticity modulates affective and social trait inferences from voices. *Philosophical Transactions of the Royal Society B: Biological Sciences, 376*(1840), Article 20200402. https://doi.org/10.1098/rstb.2020.0402

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., Fink, B., DeBruine, L. M., Jones, B. C., & Feinberg, D. R. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour, 95*, 89-99. https://doi.org/10.1016/j.anbehav.2014.06.011

Pisanski, K., Jones, B. C., Fink, B., O'Connor, J. J. M., DeBruine, L. M., Röder, S., & Feinberg, D. R. (2016). Voice parameters predict sex-specific body morphology in men and women. *Animal Behaviour, 112*, 13-22. https://doi.org/10.1016/j.anbehav.2015.11.008

Planned Parenthood. (2022a). *All About Sex, Gender, and Gender Identity*. Retrieved February 15, 2022, from https://www.plannedparenthood.org/learn/teens/all-about-sex-gender-and-gender-identity

Planned Parenthood. (2022b). *What do transgender and cisgender mean?* Retrieved February 15, 2022, from https://www.plannedparenthood.org/learn/teens/all-about-sex-gender-and-gender-identity/what-do-transgender-and-cisgender-mean

Quam, C., & Swingley, D. (2012). Development in Children's Interpretation of Pitch Cues to Emotions. *Child Development, 83*(1), 236-250. https://doi.org/10.1111/j.1467-8624.2011.01700.x

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. (Version 4.1.1) R Foundation for Statistical Computing. https://www.R-project.org/

Rathus, S. A. (2017). *Childhood & adolescence: voyages in development* (6th ed.). Wadsworth Cengage Learning.

Reed, A. E., & Carstensen, L. L. (2012). The theory behind the age-related positivity effect. *Frontiers in Psychology, 3*, Article 339. https://doi.org/10.3389/fpsyg.2012.00339

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*. (Version 2.1.6) https://CRAN.R-project.org/package=psych

Rezlescu, C., Penton, T., Walsh, V., Tsujimura, H., Scott, S. K., & Banissy, M. J. (2015). Dominant Voices and Attractive Faces: The Contribution of Visual and Auditory Information to Integrated Person Impressions. *Journal of Nonverbal Behavior, 39*(4), 355-370. https://doi.org/10.1007/s10919-015-0214-8

Rigoulot, S., Wassiliwizky, E., & Pell, M. D. (2013). Feeling backwards? How temporal order in speech affects the time course of vocal emotion recognition. *Frontiers in Psychology, 4*(367), 1-14. https://doi.org/10.3389/fpsyg.2013.00367

Rivière, E., & Champagne-Lavau, M. (2020). Which contextual and sociocultural information predict irony perception? *Discourse Processes, 57*(3), 259-277. https://doi.org/10.1080/0163853X.2019.1637204

RStudio Team. (2016). *RStudio: Integrated Development Environment for R*. (Version 1.4.1717 "Juliet Rose") RStudio, Inc. http://www.rstudio.com/

Ruffman, T., Henry, J. D., Livingstone, V., & Phillips, L. H. (2008). A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. *Neuroscience & Biobehavioral Reviews, 32*(4), 863-881. https://doi.org/10.1016/j.neubiorev.2008.01.001

Sakuta, Y., Kanazawa, S., & Yamaguchi, M. K. (2018). Infants prefer a trustworthy person: An early sign of social cognition in infants. *PLoS ONE, 13*(9), Article e0203541. https://doi.org/10.1371/journal.pone.0203541

Santrock, J. W. (2020). *A topical approach to life-span development* (10th ed.). McGraw-Hill Education.

Sauerbrei, W., & Royston, P. (2010). Continuous variables: to categorize or to model. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. International Statistical Institute. https://icots.info/icots/8/cd/pdfs/invited/ICOTS8_6D1_SAUERBREI.pdf

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. https://doi.org/10.1080/17470211003721642

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences, 107*(6), 2408-2412. https://doi.org/10.1073/pnas.0908239106

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. https://doi.org/10.1111/j.2044-835X.2012.02081.x

Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states? *Motivation and Emotion, 31*(3), 192-199. https://doi.org/10.1007/s11031-007-9065-x

Sawyer, S. M., Azzopardi, P. S., Wickremarathne, D., & Patton, G. C. (2018). The age of adolescence. *The Lancet Child & Adolescent Health, 2*(3), 223-228. https://doi.org/10.1016/S2352-4642(18)30022-1

Saxton, T. K., Caryl, P. G., & Roberts, S. C. (2006). Vocal and facial attractiveness judgments of children, adolescents and adults: The ontogeny of mate choice. *Ethology, 112*(12), 1179-1185. https://doi.org/10.1111/j.1439-0310.2006.01278.x

Saxton, T. K., DeBruine, L. M., Jones, B. C., Little, A. C., & Roberts, S. C. (2009). Face and voice attractiveness judgments change during adolescence. *Evolution and Human Behavior, 30*(6), 398-408. https://doi.org/10.1016/j.evolhumbehav.2009.06.004

Saxton, T. K., DeBruine, L. M., Jones, B. C., Little, A. C., & Roberts, S. C. (2013). Voice pitch preferences of adolescents: Do changes across time indicate a shift towards potentially adaptive adult-like preferences? *Personality and Individual Differences, 55*(2), 90-94. https://doi.org/10.1016/j.paid.2013.02.009

Saxton, T. K., Kohoutova, D., Roberts, S. C., Jones, B. C., DeBruine, L. M., & Havlicek, J. (2010). Age, puberty and attractiveness judgments in adolescents. *Personality and Individual Differences, 49*(8), 857-862. https://doi.org/10.1016/j.paid.2010.07.016

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion, 23*(7), 1307-1351. https://doi.org/10.1080/02699930902928969

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76-92. https://doi.org/10.1177/0022022101032001009

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution, 11*(9), 1141-1152. https://doi.org/10.1111/2041-210X.13434

Schieman, S. (2010). The Sociological Study of Anger: Basic Social Patterns and Contexts. In M. Potegal, G. Stemmler, & C. Spielberger (Eds.), *International Handbook of Anger: Constituent and Concomitant Biological, Psychological, and Social Processes* (pp. 329-347). Springer New York. https://doi.org/10.1007/978-0-387-89676-2_19

Schirmer, A., Feng, Y., Sen, A., & Penney, T. B. (2019). Angry, old, male – and trustworthy? How expressive and person voice characteristics shape listener trust. *PLoS ONE, 14*(1), Article e0210555. https://doi.org/10.1371/journal.pone.0210555

Schirmer, A., Simpson, E., & Escoffier, N. (2007). Listen up! Processing of intensity change differs for vocal and nonvocal sounds. *Brain Research, 1176*, 103-112. https://doi.org/10.1016/j.brainres.2007.08.008

Schroeder, J., & Epley, N. (2015). The Sound of Intellect: Speech Reveals a Thoughtful Mind, Increasing a Job Candidate's Appeal [Article]. *Psychological Science, 26*(6), 877-891. https://doi.org/10.1177/0956797615572906

Schvartz, K. C., & Chatterjee, M. (2012). Gender Identification in Younger and Older Adults: Use of Spectral and Temporal Cues in Noise-Vocoded Speech. *Ear and Hearing, 33*(3), 411-420. https://doi.org/10.1097/AUD.0b013e31823d78dc

Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zäske, R. (2014). Speaker perception. *WIREs Cognitive Science, 5*, 15-25. https://doi.org/10.1002/wcs.1261

Sell, A., Bryant, G. A., Cosmides, L., Tooby, J., Sznycer, D., von Rueden, C., Krauss, A., & Gurven, M. (2010). Adaptations in humans for assessing physical strength from the voice. *Proceedings of the Royal Society B-Biological Sciences, 277*(1699), 3509-3518. https://doi.org/10.1098/rspb.2010.0769

Sen, A., Isaacowitz, D., & Schirmer, A. (2018). Age differences in vocal emotion perception: on the role of speaker age and listener sex. *Cognition & Emotion, 32*(6), 1189-1204. https://doi.org/10.1080/02699931.2017.139339

Shaffer, D. R., & Kipp, K. (2014). *Developmental psychology: Childhood and adolescence* (9th ed.). Wadsworth Cengage Learning.

Shariff, A. F., & Tracy, J. L. (2011). What Are Emotion Expressions For? *Current Directions in Psychological Science, 20*(6), 395-399. https://doi.org/10.1177/0963721411424739

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Silva, J. M. (2016). High Hopes and Hidden Inequalities: How Social Class Shapes Pathways to Adulthood. *Emerging Adulthood, 4*(4), 239-241. https://doi.org/10.1177/2167696815620965

Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science, 16*(5), 264-268. https://doi.org/10.1111/j.1467-8721.2007.00517.x

SSEA. (2014). *About the Society for the Study of Emerging Adulthood (SSEA)*. Retrieved December 16, 2021, from http://www.ssea.org/about/index.htm

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: ambient images generate a three-dimensional model. *Cognition, 127*(1), 105-118. https://doi.org/10.1016/j.cognition.2012.12.001

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters. *Procedia Technology, 9*, 1112-1122. https://doi.org/10.1016/j.protcy.2013.12.124

Thompson, A. E., & Voyer, D. (2014). Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis. *Cognition and Emotion, 28*(7), 1164-1195. https://doi.org/10.1080/02699931.2013.875889

Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior, 33*(3), 210-216. https://doi.org/10.1016/j.evolhumbehav.2011.09.004

Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Year in Cognitive Neuroscience 2008, 1124*, 208-224. https://doi.org/10.1196/annals.1440.012

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*(12), 455-460. https://doi.org/10.1016/j.tics.2008.10.001

Tolland, L., & Evans, J. (2019, February 21). *What is the difference between sex and gender?* Retrieved February 15, 2022, from https://www.ons.gov.uk/economy/environmentalaccounts/articles/whatisthedifferencebetweensexandgender/2019-02-21

Tonks, J., Williams, W. H., Frampton, I., Yates, P., & Slater, A. (2007). Assessing emotion recognition in 9-15-years olds: Preliminary analysis of abilities in reading emotion from faces, voices and eyes. *Brain Injury, 21*(6), 623-629. https://doi.org/10.1080/02699050701426865

United Nations. (2021). *FAQ - What does the UN mean by 'youth' and how does this definition differ from that given to children?* Retrieved 16/12/2021,

from https://www.un.org/development/desa/youth/what-we-do/faq.html

United Nations. (n.d.). *FAQ - What does the UN mean by 'youth' and how does this definition differ from that given to children?* Retrieved February 6, 2022, from https://www.un.org/development/desa/youth/what-we-do/faq.html

Voyer, D., Thibodeau, S.-H., & Delong, B. J. (2016). Context, Contrast, and Tone of Voice in Auditory Sarcasm Perception. *Journal Of Psycholinguistic Research, 45*(1), 29-53. https://doi.org/10.1007/s10936-014-9323-5

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior, 17*(1), 3-28. https://doi.org/10.1007/BF00987006

Waxer, M., & Morton, J. B. (2011). Children's Judgments of Emotion From Conflicting Cues in Speech: Why 6-Year-Olds Are So Inflexible. *Child Development, 82*(5), 1648-1660. https://doi.org/10.1111/j.1467-8624.2011.01624.x

Westfall, J. (2016). *PANGEA: Power ANalysis for GEneral Anova designs* [Unpublished Manuscript]. University of Texas at Austin. http://jakewestfall.org/publications/pangea.pdf

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020-2045. https://doi.org/10.1037/xge0000014

WHO. (2021, November 17). *Adolescent mental health*. Retrieved February 6, 2022, from https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

WHO. (2022a). *Adolescent Health*. Retrieved February 6, 2022, from https://www.who.int/southeastasia/health-topics/adolescent-health

WHO. (2022b). *Gender and Health*. Retrieved February 15, 2022, from https://www.who.int/health-topics/gender

WHO. (2022c, May 20). *World Health Statistics 2022*. Retrieved July 28, 2022, from https://www.who.int/news/item/20-05-2022-world-health-statistics-2022

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pederson, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), Article 1686. https://doi.org/10.21105/joss.01686

Willis, J., & Todorov, A. (2006). First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science, 17*(7), 592-598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R* (First ed.). Rudledge. https://doi.org/10.4324/9781315165547

Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and Voice Perception: Understanding Commonalities and Differences. *Trends in Cognitive Sciences, 24*(5), 398-410. https://doi.org/10.1016/j.tics.2020.02.001

Zäske, R., Skuk, V. G., Golle, J., & Schweinberger, S. R. (2020). The Jena Speaker Set (JESS)—A database of voice stimuli from unfamiliar young and old adult speakers. *Behavior research methods, 52*(3), 990-1007. https://doi.org/10.3758/s13428-019-01296-0

Zebrowitz-McArthur, L., & Montepare, J. M. (1989). Contributions of a babyface and a childlike voice to impressions of moving and talking faces. *Journal of Nonverbal Behavior, 13*(3), 189-203. https://doi.org/10.1007/BF00987049

Zebrowitz, L. A., Boshyan, J., Ward, N., Gutchess, A., & Hadjikhani, N. (2017). The Older Adult Positivity Effect in Evaluations of Trustworthiness: Emotion Regulation or Cognitive Capacity? *PLoS ONE, 12*(1), Article e0169823. https://doi.org/10.1371/journal.pone.0169823

Zebrowitz, L. A., & Collins, M. A. (1997). Accurate Social Perception at Zero Acquaintance: The Affordances of a Gibsonian Approach. *Personality and Social Psychology Review, 1*(3), 204-223. https://doi.org/10.1207/s15327957pspr0103_2

Zebrowitz, L. A., & Montepare, J. M. (2008). Social Psychological Face Perception: Why Appearance Matters. *Social And Personality Psychology Compass, 2*(3), 1497-1517. https://doi.org/10.1111/j.1751-9004.2008.00109.x

Zupan, B. (2015). Recognition of high and low intensity facial and vocal expressions of emotion by children and adults. *Journal of Social Sciences and Humanities, 1*(4), 332-344. http://files.aiscience.org/journal/article/html/70320049.html