Emotional Speech Processing in Infants and Adults:
A Behavioral and Electrophysiological Investigation

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Chieh Kao

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Yang Zhang, Ph.D.

August 2021

Acknowledgement

I would like to thank my advisor, Yang Zhang, for his time, effort, and great support over the years. You never turn down my ideas, and you always have confidence in me even when I have little in myself. You taught me how to deal with pushbacks and be persistent. I am thankful for all the great things I learned from you.

I would also like to thank my committee members, Lizbeth Finestack, Maria Sera, and Matthew Winn, for their guidance and encouragement. Thank you, Maria, for collaborating with me on interdisciplinary research projects and helping with the participant recruitment. Thank you, Matt, for teaching me to write Praat scripts for better acoustic analyses. Thank you, Liza, for your insights into the connections between basic and clinical research and your great support as the Director of Graduate Studies.

I would also like to acknowledge Zhang lab members for their assistance and friendship: Luodi Yu and Tess Koerner, and a group of brilliant undergraduate research assistants: Jessica Tichy, Kailie McGuigan, Natasha Stark, Shannon Hofer-Pottala, Emily Krattley, Hayley Levenhagen, Megan Peterson, and Corrin Murray. A special thanks to Dr. Megha Sundara for the technical support in setting up the infant laboratory.

Finally, I would like to thank my parents, grandparents, sister, and friends for their endless love, acceptance, and support. Thank you, mom and dad, for always being there for me and listening to all my worries even if they are so small compared to the world. I could not have completed the dissertation without you.

This dissertation project was made possible by funding from the Graduate Research Partnership Program (GRPP), BryngBryngelson Research Fund, Interdisciplinary Doctoral Fellowship, Doctoral Dissertation Fellowship, and the Taiwanese Government Scholarship to Study Abroad. Research projects were also supported in part by the Brain Imaging Research Project to Yang Zhang.

Abstract

Emotional prosody is integral to successful communication as it conveys the speaker's emotional state and shapes the meaning of the words and sentences. Timely processing of vocal emotion facilitates speech comprehension as well as social interactions. However, very few studies have examined how infants process emotional speech prosody and how this ability develops from infancy to adulthood.

The current dissertation includes three original studies to address emotional speech processing from a developmental perspective. The first study aimed to characterize 3-12-month-old infants' listening attention to basic emotional prosodies in spoken words—happy, angry, sad, and neutral and the potential age and sex effects. Infants' preferential looking times showed that they listened longer to the affective than the neutral voices, especially the happy and sad speech. Significant interaction effects were observed between emotion category and acoustic parameters of vocal emotion, but there were no main effects of age and sex. The second study employed a roving multi-feature oddball paradigm to record infants' neurophysiological responses to these three basic affective prosodies against the neutral one. Infants showed distinct mismatch responses (MMRs) to different emotions in both early (100-200 ms) and late (300-500 ms) time windows, indicating their ability to extract affective speech patterns and detect emotional prosody changes at the pre-attentive level. Age- and sex-related effects were observed in the MMR data, indicating a higher degree of sensitivity of the electrophysiological measures over the behavioral measures in the first study. In the

third study, adult listeners completed the same emotional multi-feature oddball experiment. The adults showed a stronger mismatch negativity (MMN) to angry prosody and a stronger P3a to happy prosody. Gender differences continued to be observed in the adult MMN and P3a data outside attentional focus on the emotional prosody changes in spoken words. Together, the current dissertation provides empirical data on emotional processing in speech from a developmental perspective, and it has strong implications for future studies to address links between early socio-emotional development and language development in normal and clinical populations.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

I. Overview

Affect and cognition have been viewed as separate faculties and treated as independent research topics until the last 40 years (Forgas, 2008, 2012; Heider, 2013; Hilgard, 1980). With the rise of social cognitive neuroscience, affect has been recognized as a faculty of evolutionary values that plays an essential role in cognitive processing (Cosmides & Tooby, 2000; Damasio, 2006; Davidson, 2000; Isen, 1987; Neisser & Hyman, 2000; Zajonc, 1980). Cognitive processing would be vague and abstract without being projected through language, and the two co-evolve in the human mind (Barsalou, 1999; Fauconnier & Turner, 2008; Perlovsky, 2009). The inextricable link between cognition and language suggests that the affective influence should also be considered in the context of language (Barrett et al., 2007). Research on the interplay between affect and language (including both spoken and written language) emerged even more recently, and the empirical evidence is still insufficient to define their links, functions, and neural mechanisms (Hohenberger, 2011). Furthermore, the fundamental question of how humans master the extraction of affect- and language-related information from the environmental inputs remains unclear.

This dissertation aims to provide empirical evidence for the development of affective processing in the context of spoken language. By investigating infants and adults' processing of affective speech through behavioral and neurophysiological measurements, the findings can address the developmental changes of listeners' attention orientation to affective features in speech and provide the relevant neural correlates unfolding at the time scale of milliseconds.

The terms *affect* and *emotion* are used interchangeably, but there are still distinctions between the two (Shouse, 2005). *Affect* is an individual's core state and raw experience based on the neurophysiological state at the moment. Affect drives individuals' thinking and actions (Thayer, 1989), but it is abstract without being realized in language. *Emotion*, on the other hand, is the projection of an individual's affective state and can be labeled by language. Furthermore, emotion can be genuine or posed, and therefore it is both intra- and inter-personal. Even though the definitions and distinctions between *affect* and *emotion* are still under debate (Russell, 2003; Watson & Tellegen, 1985), most empirical studies examine individuals' processing of *emotion* by using stimuli of displayed expressions. Hence, the term *emotion* is adopted in the current dissertation to characterize the broadcasted emotional information in human speech. The terms *affect*, *affection*, or *affective* may still appear infrequently throughout the dissertation to alternate with *emotion* or *emotional*, but both refer to the non-abstract emotional information in human communication that serves a social role.

Emotion is crucial in human communication to express ones' internal states, address problems, and request proper responses from others. For instance, happiness and joy are contagious and important for strengthening social bonds; anger indicates the necessity of changing the current situation and is powerful in regulating social interactions; sadness signals personal loss and a need of appropriate reactions from the people around (Frijda, 2000). These social functions underline the significant role of emotions in building relationships and maintaining interpersonal interactions (Fischer & Manstead, 2008). Consequently, failure to process emotional information is viewed as a

concern in communication and even in personal well-being (Salisch, 2001; Williams, 2002). Infants and younger children are especially vulnerable since difficulties in processing socio-emotional information may negatively impact their future cognitive and language development (Fox & Calkins, 2003). While some studies have demonstrated links between emotional processing and language development (Bhullar, 2008; Singh et al., 2004; Singh, 2008), more empirical evidence is needed to delineate the developmental changes in infancy and better characterize typically developing infants' and adults' emotional processing in the context of speech and language.

II. Emotional Prosody—the Acoustic Correlates and Stimulus Selection

Emotional prosody is the emotional voice expressed through specific sets of acoustic variables in the speaker's moment-by-moment vocalizations. This dynamic signal can be decomposed into several fundamental acoustic parameters (Banse & Scherer, 1996; Hammerschmidt & Jürgens, 2007). The fundamental frequency (F0) and vocal energy (amplitudes) are the two major constituents of emotional prosody. Other acoustic parameters include but are not limited to the temporal cues (pausing, speech rate), spectral information, locations of the formants, and measurements of voice quality. The field of the perception of emotional prosody has not reached a consensus of discrete sets of acoustic combinations for each emotion due to the high acoustic variations in intra- and inter-personal production of emotional voices (Schröder, 2001; Tato et al., 2002). Because emotions can be mixed to different degrees over the arousal/valence spectrum also adds complexity to defining the specific acoustic profiles for each emotional category (Ladd et al., 1985; Scherer, 1984, 1986).

Despite the above challenges, some common acoustic variables are frequently reported in acoustic analysis studies. For instance, F0-related measures (mean, range, variability, contour), intensity-related measures (mean, range, variability, contour), and speech rate (or word duration) were frequently included in theoretical-testing or empirical experimental reports (Banse & Scherer, 1996; Jaywant & Pell, 2012; Liu & Pell, 2012; Murray & Arnott, 1993; Williams & Stevens, 1972). Basic emotional prosodies have easy-to-recognize (specific) acoustic profiles (Banse & Scherer, 1996; Johnstone & Scherer, 2000). For instance, happy prosody is usually with higher F0-related and intensity-related measures (mean, variability, range), and it is also with more energy at a higher-frequency range and moderate speech rate. Angry prosody usually has higher F0 variability, higher intensity-related measures (mean, variability, range), more energy at higher-frequency range, and faster speech rate. Sad prosody is usually with lower F0-related and intensity-related measures (mean, variability, range), more spectral noise, and a slower speech rate. The intrinsic acoustic constituents of these emotional prosodies may influence the subjective experience of the emotional signals. Due to infants' limited autobiographic experiences of verbalizable emotional feelings (Ekman, 1984; Shouse, 2005), their attention to the emotional prosody may be directed by the acoustic features as well. Therefore, including these acoustic variables in investigating infants' perception of emotional speech can better address their attention orientation to different emotional portrayals in human voices.

Previous studies on emotional speech recognition used acted, elicited, or natural emotional prosodies as the stimuli (Koolagudi & Rao, 2012; Swain et al., 2018). Acted

emotional speech stimuli are recordings from professional voice performers. They are asked to express the target vocal emotions with neutral linguistic materials, with or without the given emotional context, to facilitate their acting. Most empirical reports used the acted emotional speech because the stimuli can be generated consistently and effectively in a more controlled condition. Elicited emotional speech stimuli are usually collected during conversations of assigned emotional context in a laboratory. Performers do not know the target emotional situation but may be aware that they are recorded. The elicited emotional speech may be more natural than the acted ones, but it still differs from the day-to-day emotional expressions in natural conversations. Unlike acted or elicited emotions, natural emotional speech is hard to collect. Researchers may obtain real-life vocal emotional expressions in telephone recordings or public conversations, but there are ethical and legal constraints such as privacy and copyright. Therefore, acted or elicited emotional speech may be more feasible for most empirical studies.

The current dissertation adopted acted emotional speech for three reasons: acted speech provides (1) consistent quality of the emotional prosody and high quantity of lexical items (e.g., Alpert et al., 2001; Burkhardt & Sendlmeier, 2000; Dupuis & Pichora-Fuller, 2010; Makarova & Petrushin, 2002), (2) more expressive and intense portrayals of the emotions than the elicited or natural ones (El Ayadi et al., 2011; Schröder et al., 2001; Williams & Stevens, 1972), and (3) equally valid expressions from the professional performers regardless of the genuineness of the affect (Marty, 1908; Scherer, 1986).

III. Methodologies in Developmental Studies on Emotional Speech Perception

Infants' listening attention and neural responses to the emotional speech prosody can help address the processing of auditory affective signals during their first year of life. However, methodologies in developmental sciences are limited because infants' voluntary responses are restricted to their cognitive and motor development (Rakison & Yermolayeva, 2010; Stager & Werker, 1998; Werker & Fennell, 2009). In behavioral paradigms, researchers can only measure infants' overt reactions and are limited to interpreting infants' relatively less reliable behavioral responses. With the advance of neuroimaging techniques and standardized protocols for infants, the neurophysiological data may complement the behavioral data with additional information that contribute to better understanding the developmental changes in early cognitive and affective processing. The current dissertation included both behavioral and neurophysiological measurements in studying infants' emotional speech perception. Adults' neurophysiological data were also recorded to compare the neural correlates of emotional voices in young and mature typical listeners. The following sections review the two experimental protocols used in the dissertation.

A. Central Fixation Paradigm

The central fixation paradigm, or so-called look-to-listen paradigm, is a testing protocol that uses infants' looking times to infer their attention to auditory stimuli. The link between infants' looking times and attention to the auditory stimuli was first established in an earlier experimental design called the *intermodal preferential listening paradigm* (IPLP, Golinkoff et al., 1987; Hirsh-Pasek & Golinkoff, 1996). In the IPLP,

infants are presented with two side-by-side images and one audio track that matches one of the images. Infants tend to look longer to the matched image, indicating that their visual fixation may reflect their selective attention to the auditory inputs. This use of IPLP allows developmental scientists to investigate preverbal infants' cognitive processing through their natural gaze patterns (Golinkoff et al., 2013), especially when examining infants' emerging speech-and-language ability.

In less than 35 years since the creation of IPLP, modified versions of this experimental protocol were introduced (Fernald et al., 2008). One of them is the *central fixation paradigm* (or so-called look-to-listen paradigm, or sequential preferential listening paradigm; Shultz & Vouloumanos, 2010). In a central fixation paradigm, experimenters present one static image (not related to the audio file, e.g., a checkerboard) along with the auditory stimuli. The visual and auditory inputs are both presented in front of the infant, sharing the same source. If the infant is interested in the sound, they tend to look at the sound source (i.e., the image). When infants are no longer interested in the sound, they would look away, and the experimenters would terminate the sound. Each type of stimuli would be repeated several times, and the experimenters can compare infants' accumulative looking times to each stimulus. Differential looking times indicate that the infant represents the auditory stimuli differently, with longer listening times associated with more attention to the sound (Haith, 1980). In Shultz and Vouloumanos' report (2010), 3-month-old infants can already learn the look-to-listen contingency. Therefore, the central fixation paradigm is appropriate for presenting multiple emotional speech prosodies to infants older than three months.

B. Electroencephalography (EEG)

Electroencephalography (EEG) measures the high-dimensional time-series data of a participant's post-synaptic electrical signals from a collective of neurons (Luck, 2014). During an auditory task, the brain regions that respond to the auditory information are activated. The sum of the collective post-synaptic electrical signals relative to the onset of the sound is called evoked event-related potential (ERP). Since the neurophysiological signals usually contain task-irrelevant noise, a grand mean ERP was taken from averaging hundreds of ERP waveforms across trials to reveal the meaningful ERP signals from the EEG background noise. The task-relevant ERP signals in the average waveform are the observable peaks of positive or negative deflections called ERP components. ERP components are usually labeled by the latency (millisecond relative to the sound onset) and polarity (positive or negative). For instance, the auditory N1 response in adults is the negative deflection around 100 ms after the sound onset, and P2 is the positive deflection around 200 ms after the sound onset. To compare listeners' ERPs to different stimuli, amplitudes of the ERP component (in micro-voltage) are used as the dependent variable. The ERP components reflect the neural processing of the auditory inputs unfolding at the millisecond level, and the elicitation of the ERP component does not necessarily require listeners' active engagement in the task. Therefore, the current dissertation adopted the EEG method to record both infants' and adults' neural activities to emotional speech. The target ERP components—mismatch negativity (MMN) and P3a—are reviewed in the following sections.

Mismatch Negativity (MMN) and Mismatch Response (MMR)

The auditory mismatch negativity (MMN) is an ERP component that registers the change of the sound in a passive listening task using the oddball paradigm (Näätänen et al., 1978; Näätänen et al., 2007). In this passive listening task, the participant was asked to ignore the continuous stream of sounds and complete a quiet task-irrelevant activity (e.g., watching a silent movie, reading a book). There are two types of auditory stimuli in the oddball task—a frequent one (*Standard*, typically presented over 85% of the time) and an infrequent one (*Deviant*, typically presented less than 15% of the time). The frequent *Standard* sound establishes a sensory memory trace in the listener's central auditory system. When the stream of repeating *Standard* sound is interrupted by the *Deviant* sound, the auditory system detects the change and responds to the *Deviant* sound differently from the *Standard* sound. The degree of perceptual differences between the *Standard* and *Deviant* sounds is characterized by the difference ERP waveform, derived by subtracting the *Standard* ERP from the *Deviant* ERP. If the listener's auditory system captures the sound change, a negative deflection—the MMN—appears around 100 – 300 ms after the onset of the stimulus change in the difference ERP waveform. The MMN is usually recorded at frontal to central electrodes (i.e., electrodes Cz and Fz) over the scalp. Some reports show a left- or right-lateralized MMN response depending on the stimuli being used (e.g., Schirmer et al., 2005; Thönnessen et al., 2010). Notably, infants tend to show a positive peak or a mix-polarity response in the MMN window (see the comparison table in He et al., 2007). Therefore, researchers usually use MMR (mismatch response) to address this infant-version MMN

component. Because the MMN/R is recorded in a passive listening context, it is thought to index a listener's automatic processing and pre-attentive sensitivities of the central auditory system. Although the latency of MMN is relatively short, it is elicited by the perceptual, not merely acoustic, differences and involves some degrees of cognitive processing (Horváth et al., 2008; Näätänen et al., 2007).

The P3a Response

The P3a component reflects the involuntary attention allocation to the unexpected auditory event (Escera et al., 2000; Friedman et al., 2001; Polich, 2007). Similar to the MMN, the P3a can also be elicited in an oddball paradigm. The P3a is a positive deflection in the difference ERP waveform (*Standard* ERP subtracted from the *Deviant* ERP) around 300 – 500 ms after the onset of the sound change. It is usually observed at centro-frontal electrode regions (e.g., electrodes Cz and Fz), and its peak amplitudes increase from the frontal to parietal electrode sites (Polich, 2007). The P3a response occurs after the MMN, and it indexes an update of the auditory context with a novel auditory event after evaluating and comparing the *Standard* and *Deviant* stimuli. This relatively late ERP component may mark the categorization of the stimulus (Gentili et al., 2014) and is investigated in previous EEG studies on the processing of emotional prosody (e.g., Pakarinen et al., 2014; Thierry & Roberts, 2007; Zora et al., 2020). Because of the temporal proximity between P3a and the preceding negative MMR, it is challenging to record a clean P3a response from infants. A newborn study successfully recorded both the early negative MMR and the P3a by short (100 ms) non-speech *Standard* and *Deviant* sounds, and the P3a response was attributed to both the novelty

and acoustic energy of the sounds (Kushnerenko et al., 2007). Similar sets of short tones also elicited the P3a component, sometimes so strong that it masks the preceding negative MMR, in infants before the age of one (Kushnerenko et al, 2002). However, the "P3a" elicited by longer speech stimuli (500 – 1000 ms) is still called a mismatch response (e.g., slow-positive MMR, late MMR), because the speech-elicited discriminatory ERP may appear as a broad positive deflection covering the windows for both MMN and P3a (Peter et al., 2016). Some researchers believe that this late positive MMR is functionally similar to the adult MMN, indexing an automatic auditory change detection (Leppänen et al., 1999; Pinko et al., 1999); and other researchers believe that it is the infant analogue of P3a that indexes the involuntary attention switch to the *Deviant* sound change (Alho et al., 1990; Trainor et al., 2001).

Multi-Feature Oddball Paradigm for MMN and P3a

The classic oddball paradigm is limited to testing listeners' pre-attentive neural sensitivities to two types of auditory stimuli. If the research questions involve more than two sound categories, the participant needs to complete several oddball tasks to test every sound contrast (e.g., Carminati et al., 2018). Because one oddball task may take at least 25 minutes to complete, it is not feasible to ask wide-awake infants to complete several oddball tasks and record high-quality EEG signals. With increasing demands of using this passive listening paradigm in younger or clinical populations, a modified oddball paradigm that can examine multiple *Deviants* against the *Standard* was introduced—the multi-feature oddball paradigm (or optimal design, Näätänen et al., 2004; Pakarinen et al., 2007). In a multi-feature oddball paradigm, the frequent *Standard*

sound is presented 50% of the time, and the multiple *Deviant* sounds equally share the rest of the 50% presentation. This paradigm may be more demanding on the central auditory system, because a smaller proportion of the *Standard* trials make it harder to establish a stable auditory memory trace for later novel sound detection. Consequently, the MMN and P3a elicited in the multi-feature oddball paradigm tend to be weaker than those elicited in the classic oddball paradigm. Despite this downside, the advantage of time efficiency of using this paradigm makes it popular among studies testing infants and young children. Furthermore, two newborn studies have successfully recorded sleeping newborns' differential ERP response to one *Standard* and eight *Deviant* sounds (Kostilainen et al., 2020; Kostilainen et al., 2018). The current dissertation only examined three emotional prosodies (three *Deviants*) against the neutral prosody (one *Standard*) in older infants and adults, so the multi-feature oddball paradigm was adequate in recording the ERPs to three *Deviants* in a single EEG recording session.

IV. Planned Studies and Research Questions

The role of socio-emotional information in speech and language development has not been fully defined due to the lack of empirical data especially from infants before the age of one. With the central fixation paradigm, infants' looking times can be used as a proxy of their selective attention to different emotional speech sounds. For emotional voices that draw similar listening attention, the multi-feature oddball paradigm with EEG recordings can further determine if infants' central auditory systems already distinguish the emotional prosodies that are readily available in their differential behavioral responses. By measuring adult listeners' neural activities to the emotional

speech with the same multi-feature oddball task, a mature pre-attentive neural sensitivity pattern can be obtained and compared to the infants' neurophysiological data. Together, the findings can address listeners' attention to affective speech as a function of their listening experiences and add empirical data on the unfolding of affective information in the pre-attentive neural system.

A. Study 1: Emotional Speech Processing in 3- to 12-Month-Old Infants: Influences of Emotion Categories and Acoustic Parameters

Research questions:

1. What are the developmental changes in infants' listening attention to happy, angry, sad, and neutral prosody?
2. Is biological sex a potential factor in infants' emotional speech processing?
3. What are the roles of the acoustic variables (fundamental frequency, word duration, intensity variation, harmonics-to-noise ratio, and spectral centroid) in infants' differential listening attention to emotional prosody?

Expected outcomes:

1. Because infants have been found to pay more attention to the relatively positive emotion (Singh et al., 2002) and sounds with a higher pitch (Fernald & Kuhl, 1987), they will listen the longest to the happy prosody but the shortest to the sad prosody. Older infants may show differential attention to happy and angry prosodies (both high-arousal emotions), but not younger infants.
2. There may be biologically-based sex differences in male and female infants in their selective attention to emotional speech.

3. Previous literature showed that infants generally listened more to speech with a higher pitch (Fernald & Kuhl, 1987), more energies in higher frequencies (Cooper & Aslin, 1994), and longer word duration (Fernald & Simon, 1984). Even though it is unclear how infants pay attention to intensity variations and harmonics-to-noise ratio, the acoustic variables are expected to influence infants' looking time to the emotional speech. As no previous infant studies have systematically investigated the role of the acoustic parameters in emotional speech perception, it remains an exploratory question how acoustic features mediate infants' responses to each vocal emotion.

B. Study 2: Infants' Neural Sensitivity to Emotional Prosody Differences in Spoken Words

Research questions:

1. Can infants categorize emotional prosody over non-repeating words by showing distinct MMR to emotional prosody change from neutral to happy, angry, and sad?
2. What are the developmental changes in infants' neural sensitivities (i.e., MMR amplitudes) to the emotional prosodies?
3. Is there a sex effect in young infants' pre-attentive neural sensitivities to emotional prosody?

Expected outcomes:

1. Infants' central auditory system is expected to be able to extract the emotional prosodic categories over varying linguistic items and show distinct MMRs to different emotional voices.

2. In line with the developmental literature, the older infants will show more adult-like MMR (i.e., more negative-going) than the younger infants (Cheour, 2007). Whether or not this developmental trend is consistent across all the emotional prosodies remains an exploratory question.
3. If sex differences in emotional information processing are mainly attributable to learned social factors, biological sex might not have an effect on infants' MMRs to the emotional speech.

C. Study 3: Gender Differences[1] Revealed Outside the Focus of Attention to Emotional Prosody Variation in Spoken Words

Research questions:

1. Can the current study replicate previous findings of listeners' negative bias (stronger MMN to anger in voices) using a multi-feature oddball paradigm with non-repeating spoken words?
2. How does the processing of happy, angry, and sad prosodies unfold in the pre-attentive neural system in terms of the emotion-modulated MMN and P3a amplitudes?
3. Is there a sex/gender effect in adult listeners' pre-attentive neural sensitivities to emotional prosody? If so, what do the gender differences in the neural correlates of emotional speech imply men and women's differential behavioral reactions to the emotional information?

[1] Sex/Gender differences were used interchangeably in early and even some of the recent literature on emotional processing. In sociolinguistics, sex is a biological category that is binary, and gender is a socially constructed category (Eckert, 1989). The differences between the terms *sex* and *gender* are not the focus of the current dissertation. The term "sex differences/effect" is used in the first two infant studies, and the term "gender differences/effect" is used in the third adult study.

Expected outcomes:

1. Based on the literature, adult listeners will show stronger MMN responses to angry prosody even if the presentation of constantly changing spoken words may tax the central auditory processing.
2. The P3a component is subsequent to MMN and registers more cognitively involved sound evaluation and appraisal. As such, the angry prosody may not necessarily elicit the strongest response in this later neural discriminatory stage as it does in the MMN window.
3. Consistent with previous adult behavioral studies, male and female listeners will show different MMN and P3a components to the emotional speech. In particular, female listeners may show stronger MMN and P3a amplitudes.

Chapter 2: Emotional Speech Processing in 3- to 12-Month-Old Infants: Influences of Emotion Categories and Acoustic Parameters (Study 1)

This chapter has been submitted to the Journal of Speech, Language and Hearing Research. It has been revised based on the reviewers' comments and is currently under review.

I. Introduction

Language development takes place in a socio-emotional environment that includes both linguistic and social inputs (Chong et al., 2003; Conboy et al., 2015; Golinkoff et al., 2015; Ramírez-Esparza et al., 2014). One source of important social information in natural speech is emotional prosody, the way that people express different emotions with their voices. Emotional prosody plays a major role in infants' early interaction with caregivers. Young infants with limited lexical skills rely on vocal emotions to communicate, share affection, and play with their conversational partners (Walker-Andrews, 2008). Reciprocally, caregivers make use of emotions in voice to guide and regulate infants' behaviors in uncertain or even dangerous situations (Vaish & Striano, 2004). For these reasons, differentiating and understanding emotional information in speech is indispensable to infants' socio-emotional and communicative skills. Yet, very little is known about the early development of emotional speech processing in the first year of life.

Emotional prosody is not only important for infants' concurrent communication but also central to their future language and cognitive development (Feldman Barrett et al., 2017; Hoemann et al., 2019; Hohenberger, 2011). Some recent empirical works pointed to the link between emotional speech and early language learning, but noting

that emotional contexts are not always facilitative. For instance, 7.5-month-old infants cannot recognize the words they learned in a different emotional tone (Singh et al., 2004). A follow-up study further showed that young infants might prioritize the affective cue over phonemic cue and falsely recognize similar-sounding non-words with the same emotional tone, but not correctly recognizing the target word with a different emotional tone (Singh, 2008). To refocus infants' attention to the crucial phonemic cues to learn new words, Singh (2008) introduced multiple emotional tones to create an enriched word-learning context. With high emotional prosodic variations, 7.5-month-old infants successfully recognized words presented in a novel emotional voice. This ability to generalize the learned phonemic cues across paralinguistic contexts was only previously observed in 10 month-old infants, who may better leverage the affective cues in word learning (Singh et al., 2004). Older infants and children can further follow the vocal emotional cues to navigate ambiguous information (Berman et al., 2010; Paquette-Smith & Johnson, 2016). In this regard, simple affective cues with low acoustic variations may compete with the crucial phonemic cues in younger but not older infants' word learning, while introducing more emotional variants or increasing input variability (as typically found in infant-directed speech) may encourage infants to extract the invariant phonetic features and promote a more robust word representation (Apfelbaurm & McMurray, 2011; Houston, 1999; Houston & Jusczyk, 2000). Despite the prevalence and importance of emotional prosody in natural speech, developmental studies on spoken language tend to focus on phonetic and phonological processing, and infants' emotional speech perception has not been thoroughly studied (Grossmann, 2010). Furthermore,

very few infant studies directly incorporated the acoustic components of emotional voice into explaining infants' listening behaviors. One report systematically compared 6-month-old infants' selective attention to happy, sad, and neutral speech sounds with separate acoustical and looking-time analyses (Singh et al., 2002). The same report suggested that positive affect may be the main determinant of infants' listening attention, and the relevant acoustic features such as the mean fundamental frequency may be the secondary determinant. The present study followed up this idea by including the angry prosody and examined within-infant listening preference to four emotional prosodies (happy, angry, sad, and neutral). The roles of emotion-relevant acoustic parameters were also directly included in examining infants' attention to the emotional information in speech.

Acoustic Properties of Emotional Prosody in Speech

Emotional prosody in human voices is mainly registered by the mean, range, and variations of the fundamental frequency (F0, pitch of the sound) and the sound intensity level (Banse & Scherer, 1996). It is also finely characterized by other temporal and spectral acoustic parameters such as speech rate, pausing, and energy distribution in the spectrum (Bachorowski & Owren, 2008; Johnstone & Scherer, 2000; Murray & Arnott, 1993). Generally, happy and angry sounds are expressed through greater F0 measures (mean, range, and variations), greater intensity measures (mean, range, and variations), and faster speech rate (i.e., shorter word durations) (see comparison tables in Banse & Scherer, 1996; Johnstone & Scherer, 2000). On the contrary, sad voices tend to have lower or compressed F0- and intensity-related measures (mean, range, and variations),

and slower speech rate (i.e., longer word durations) (Banse & Scherer, 1996; Johnstone & Scherer, 2000).

While F0, intensity, and word duration are the key acoustic features of vocal emotions, speech quality measures such as harmonics-to-noise ratio (HNR, breathiness of the sound) and spectral centroid (brightness of the sound) also contribute to listeners' emotional speech recognition (Amorim et al., 2021; Benders, 2013; Liu & Pell, 2012). For instance, happy and sad voices have a relatively higher HNR and sound less breathy than angry voices (Liu & Pell, 2012; Patel et al., 2011), and angry voices have higher variations in HNR (Jaywant & Pell, 2012). For energy distribution along the spectrum, happy and angry voices usually have higher spectral centroids and sound brighter than sad sounds (Mokhsin et al., 2014; but also see Cunningham et al., 2018). Even with these and many more acoustic features, there is no predetermined set of acoustic parameters that can perfectly capture authentic emotional prosody (Schröder, 2001). In the current study, we adopted the top five acoustic predictors of perceived vocal emotion in a recent longitudinal study (Amorim et al., 2021) to explain infants' listening patterns. The five acoustic variables were: (1) mean fundamental frequency (F0); (2) word duration; (3) intensity variation; (4) harmonics-to-noise ratio (HNR); and (5) spectral centroid.

Infants' Responses to Basic Emotional Prosodic Categories and Developmental Changes

Infants' auditory perception of emotion has not been as thoroughly studied as the visual perception of facial expressions. Studies suggest that they are generally good at

picking up happy sounds (Grossmann, 2010). One early report found that newborns opened their eyes more when listening to their maternal language (English) in a happy voice than sad and neutral voices, but they listened similarly to happy and angry sounds (Mastropieri & Turkewitz, 1999). While it is possible that newborns were simply paying attention to the acoustic correlates of high-arousal vocal expressions (higher F0 and intensity), newborns in this study responded equally to all emotional voices in a foreign language. These results indicate that newborns already show differential listening attention to vocal expressions of emotions, and their listening patterns cannot be entirely explained by the acoustic information (Aldridge, 1994). Walker-Andrews and Grolnick (1983) examined infants' listening sensitivity to happy and sad sounds by switching the speech from one emotion to another in a habituation task. When comparing the listening times to the switched emotion, three-month-old infants showed 10-fold more increased listening times to the happy sound (when switched from sad) than the sad sound (when switched from happy). The findings demonstrated easier voice change detection from sad to happy sounds and may suggest a listening bias toward happy prosody. In the same study, five-month-old infants also detected emotional voice change in both presenting orders, but no happy prosody bias was observed. Follow-up studies used a similar testing protocol and included angry prosody for comparisons (Flom & Bahrick, 2007; Walker-Andrews & Lennon, 1991). Infants older than five months were found to detect vocal emotional change reliably from any emotional contrasts (any two emotions from happy, angry, and sad), except when the change was from angry to happy voice (Walker-Andrews & Lennon, 1991). These results suggest that infants before the age of one can

already differentiate between basic emotional prosody, with an early listening preference toward the happy voice and some degree of confusion between happy and angry prosody. There is evidence for an early developmental change as infants younger than five months were confused more when angry prosody was included in the task, but not the older infants (Flom & Bahrick, 2007). One limitation is that these findings were largely restricted to tests using binary (pair-wise) emotional change detection (except the newborn study in Mastropieri & Turkewitz, 1999). As emotional speech is much more complex than a binary contrast, there is a need to examine within-infant responses to more than two vocal emotions.

The literature also suggests a gradual change in infants' sensitivity to different emotional prosodies over their first year of life. Newborns are more responsive to happy sounds (Mastropieri & Turkewitz, 1999), demonstrating basic discrimination between happy and the other emotions. Three-month-old infants can also discriminate between happy and sad sounds, but they only succeed when the sad prosody was presented first (Walker-Andrews & Grolnick, 1983). This inconsistent discrimination of the two emotions showed that young infants' emotional prosody processing is still immature and unstable at this age. It also implies an early listening preference for the happy voice. Five-month-old infants are no longer limited by the sound presenting order and can successfully differentiate between happy, sad, and even angry vocal expressions (Walker-Andrews & Lennon, 1991), showing a more mature emotional prosody discrimination. When infants turn seven months, they can differentiate between happy and neutral sounds even when some asynchronous talking-face videos were presented

(Walker, 1982). By nine months, infants can use their parents' vocal expressions to make appropriate decisions in uncertain situations (Mumme et al., 1996; Paquette-Smith & Johnson, 2016). These studies showed that infants become more sophisticated listeners of emotional prosody as they gain more listening experiences. Even though this developmental trend has been primarily derived from sound discrimination tasks, we would expect to see older infants showed more distinct listening patterns than younger infants for the four different categories of vocal emotional expressions.

Acoustic Contributors to Attentional Processing of Emotional Prosody in Infancy

Previous studies on infants' emotional speech perception have seldom included analyses of the acoustic parameters that may help explain their listening attention (except Singh et al., 2002). Most reports focused on infants' preference for infant-directed speech (IDS) (ManyBabies Consortium, 2020) and its relevant acoustic correlates (e.g., Fernald & Kuhl, 1987), but not the emotional component within IDS and the relevant acoustic features. Singh, Morgan, and Best (2002) conducted serial experiments to investigate emotional voices (happy, sad, and neutral) independently from the speech style of IDS (baby talk, per the original report) and adult-directed speech (ADS) in 6-month-old infants. Longer listening times to happy than neutral speech were observed across speech styles, but longer listening times to neutral than sad speech were only observed when the neutral speech was in IDS (featured by a higher pitch). The authors concluded that relatively positive affect is the main determinant of infants' attention, and the acoustic feature (i.e., the mean fundamental frequency, F0) is the secondary determinant.

Due to a lack of systematic report on the roles of other acoustic parameters in infants' emotional speech processing, we hereby review some key acoustic contributors to infants' preference for infant-directed speech (IDS)—a speech style that is closely related to emotional speech. There is a consensus that infants prefer IDS to adult-directed speech (ADS) (ManyBabies Consortium, 2020). The general explanation is that infants pay more attention to the acoustic features in IDS, such as a higher mean F0 and a lengthened word duration (Fernald & Simon, 1984; Fernald et al., 1989; Stern et al., 1982). Indeed, infants listen more to speech with a higher mean F0 when the sound intensity is held constant (Fernald & Kuhl, 1987; Masapollo et al., 2016). Furthermore, the spectral information at higher frequencies is crucial in determining young infants' listening preference, as it has been shown that removing this information reduces infants' listening bias to IDS (Cooper & Aslin, 1994). As for word durations, an age-dependent listening preference has been observed. Infants younger than six months attend more to words with longer duration, whereas infants older than eight months do not (Kitamura & Notley, 2009; Panneton et al., 2006). In other words, younger infants preferred lengthened word durations as in the IDS, but not the older infants.

Past evidence on the roles of intensity variation and harmonics-to-noise ratio (HNR) in IDS is less clear than F0 and word durations. Sound intensity levels have usually been controlled in infant listening tasks. Thus, the previous studies seldom included intensity-related measures. HNR was rarely measured, for the breathy voice quality has not been the focus of infants' preferential listening. One recent study showed that IDS sounds breathier, and this breathy voice may be used to soothe or calm the

infants (Miyazawa et al., 2017). Even though the relation between breathiness in voice and infants' listening preference is indirect, HNR is worth quantifying to expand our understanding of early emotional speech perception. In summary, mean F0, spectral information, and word duration have all been shown to be related to infants' listening preference, and developmental differences may exist for the preference of word durations. Intensity variations and HNR are important acoustic constituents of emotional prosody, and they may be relevant to infants' emotional speech perception. By including these acoustic variables, we can begin to understand how acoustic components act on early listening attention to vocal expressions of emotions.

Current Study

The current study serves to fill the knowledge gap on infants' emotional prosody perception by investigating 3- to 12-month-old infants' listening attention for four basic vocal emotions—happy, sad, angry, and neutral. In addition, we included five relevant acoustic parameters—mean fundamental frequency (F0), intensity variation, word duration, harmonics-to-noise ratio (HNR), and spectral centroid—to examine their roles in infants' listening attention to emotional speech. We adopted the infant-controlled central fixation paradigm (also called the look-to-listen paradigm) used by Shultz and Vouloumanos (2010) to investigate within-infant listening attentiveness to the four emotions. In this paradigm, infants' looking time during each sound presentation was used as a proxy measure of their listening attention. In accordance with previous reports on infants' preference for the positive voice (Singh et al., 2002), we predict that the 3- to 12-month-old infants in the current study should listen longer to the happy prosody.

Angry and happy voices share similar acoustic profiles (Tato et al., 2002), and young infants tended to confuse the two (Flom & Bahrick, 2007). Therefore, we expected to see an age effect such that the older infants would show more attention to the happy voice than the angry voice, but the younger infants would listen similarly to the two emotions. Past evidence indicates that infants can discriminate sad emotions from other emotions, but very few reports directly tested infants' listening preference for sad sounds. Singh and colleagues (2002) observed a shorter listening time to the sad than the neutral voice in 6-month-old infants that may be explained by the negative affect and low-pitched nature of the sad sound, but infants younger than 6 months were not tested. If pitch plays a major role in emotional speech perception, younger infants should pay the least attention to sad sounds. Sadness in voice is also acoustically marked by longer word durations. If word duration plays a major role, younger infants would pay more attention to the sad voice that has lengthened word durations as in the IDS.

In addition to examining age differences, our study also examined sex differences in emotional prosody perception. Although one preferential listening study using IDS did not show a significant sex effect (Fernald & Simon, 1984), there is some acoustic evidence that mothers used different pitch ranges when interacting with male and female infants (Kitamura & Burnham, 2003). It is thus a legitimate question whether boys and girls process emotional prosody differently within the first year of life.

II. Method

A. Participants

The final sample for statistical reports included 43 infants between the ages of two months 26 days and 11 months 11 days (male = 22, female = 21; mean age = 7.6 months or 231 days). Initially, 46 typically developing infants from three to 12 months (male = 25, female = 21; mean age = 7.6 months or 229 days) were recruited through advertisements, word of mouth, and the infant participant pool of the Institute of Child Development at the University of Minnesota. All infants were born full-term (38 – 42 weeks), healthy with normal hearing, and from English-speaking families. The experimental protocol was approved by the local Institutional Review Board. Three infants were excluded from further analysis due to vomiting (n = 1), diaper changing (n = 1), or noise interruption (n = 1) during the experiment. Parents signed the informed consent for their children prior to the participation and received $20 as monetary compensation upon completion.

B. Materials

The speech stimuli included eighteen monosyllabic words spoken in neutral, happy, sad, and angry prosodies by a young female speaker. The words were “bar”, “base”, “chair”, “chat”, “choice”, “dog”, “germ”, “match”, “merge”, “mill”, “sail”, “shack”, “shirt”, “tool”, “turn”, “void”, “which”, and “yes”. These words were randomly selected from a phonetically balanced list (Northwestern University Auditory Test No. 6, NU-6; Tillman & Carhart, 1966). The recordings of the words in different emotional prosodies were from the Toronto Emotional Speech Set (TESS, Dupuis & Pichora-

Fuller, 2010). The sounds were sampled at 24,414 Hz, with the mean sound intensity levels equalized using Praat 6.0.40 (Boersma & Weenink, 2020). Table 1 summarizes the mean fundamental frequency (F0), duration, intensity variation, harmonics-to-noise ratio (HNR), and spectral centroid in each emotional prosody. These five acoustic measures are commonly used to characterize different vocal emotions (Amorim et al., 2021; Banse & Scherer, 1996; Johnstone & Scherer, 2000; Mani & Pätzold, 2016), and they are included in the later statistical analysis.

Table 1. The acoustic properties of each emotional prosody.

Emotions	*Mean F0 (Hz)*	*Duration (ms)*	*Intensity Variation (dB)*	*HNR (dB)*	*Spectral centroid (Hz)*
Angry	216.88	661	10.30	7.89	2160.65
Happy	223.61	756	10.19	16.68	1151.06
Sad	174.58	831	9.59	17.5	630.87
Neutral	190.13	684	7.84	17.03	850.37

Note. The averaged values of the 18 words were used to report the mean fundamental frequency (F0), word duration, intensity variation, harmonics-to-noise ratio (HNR), and spectral centroid in each emotional prosody.

We used a customized Praat script to concatenate the 18 words with the same emotional prosody into a 32-second trial, with 1-second silence between adjacent words. Four randomized word orders were created for word concatenation (see Appendix A for the four wordlists), and each word order was used for happy, angry, sad, and neutral prosodies. This gave us a total of 16 trials which were presented in a randomized block design. To familiarize the infants with the listening procedure, we also included a 32-

second music clip with piano and theremin (an electronic musical instrument) as the pretest stimulus.

C. Apparatus

The experiment was conducted in a quiet room with walls covered with thick ceiling-to-floor black curtains. The room was only lit by two dim lamps at two front corners. Infants sat on their caregivers' lap and were 55 inches away from a 22-inch LCD monitor. A video camera was placed 8 inches below the monitor to record the whole session. The stimuli were presented through Habit X (Cohen et al., 2000) on an Apple MacPro desktop computer outside the curtained-off room. The speech stimuli were presented at 55 dB SPL through two hidden speakers behind the monitor. During the task, the caregivers listened to continuous music irrelevant to the current task through circumaural headphones (Peltor series 7000). An experimenter sitting outside the curtained-off room observed by manually pressing a key on the computer keyboard to code infants' looking behaviors through the camera projected to a multifunctional computer monitor in a picture-in-picture mode. The experimenter would long-press the key "5" when the infant looked at the monitor and release the key once the infant looked away.

D. Procedure and Experimental Design

An infant-controlled central fixation paradigm (i.e., look-to-listen paradigm, Shultz & Vouloumanos, 2010) was adopted to examine infants' listening attention to happy, angry, sad, and neutral prosodies. Before each trial started, an animated ball appeared in the center of the screen to get the infant's attention. Once the infant's eye

gaze was fixated on the screen, the trial would start with playing experimental sounds and a static bright-colored checkerboard image on the screen. The infant's total looking time at the screen was monitored and recorded in each trial, and the trial would be terminated once the infant looked away for more than 2 seconds or when the 32-second sound file ended. When a trial ended, the attention-getter (the animated ball) resumed and prepared the infant for the next trial. The experiment was controlled by a trained experimenter.

The experiment was composed of one pretest and 16 test trials (four emotions each presented in four wordlists). In the pretest trial, infants listened to a 32-second music clip with piano and theremin (an electronic musical instrument) to be familiarized with the listening procedure. The order of the 16 test trials was pseudo-randomized. We first used the order of the wordlists to create four blocks, and then we randomized the four emotions within each block. An additional rule was that the same emotional prosody would not be presented consecutively. The orders of the wordlist and emotion were counterbalanced across infants. The listening test lasted 5 ~ 10 minutes.

E. Data Analysis

The looking time for each trial was calculated by offline frame-by-frame video coding (PsyCode, http://psy.ck.sissa.it/). If an infant missed a trial or the experimenter terminated a trial prematurely, the trials would be removed without any data interpolation or replacement (four trials were removed out of the total 688 trials). The trials with listening times shorter than one second (10 trials) or reaching the maximum length of the sound file (one trial) were also excluded from further analysis (Shultz &

Vouloumanos, 2010). All participants whose data were included had two or more trials for each emotion.

The acoustic variables were calculated trial-by-trial after we obtained the offline looking time of each trial for each infant. For a particular trial, we calculated the mean acoustic measures up to the last complete word that the infant heard before the trial stopped. For example, if an infant listened to a trial for 15.5 seconds, which corresponding to the middle of the 10th word in the original sound file, we averaged the mean fundamental frequency (F0), intensity variation, word duration, harmonics-to-noise ratio (HNR), and spectral centroid of the first *nine* complete words that the infant heard in this trial (i.e., this sound file) to be the five acoustic variables for this particular trial. Through this trial-by-trial acoustic analysis, the five acoustic variables can be directly included in the statistical model using trial-level looking times as the dependent variable. This acoustic analysis was completed in customized Praat and R (https://www.r-project.org/) scripts.

All statistical analyses were completed in R with the packages "lme4" (Bates et al., 2015), "lmerTest" (Kuznetsova et al., 2017), and "emmeans" (Lenth et al., 2018). We used a linear mixed-effect model to take the looking time of each individual trial as the dependent variable. The looking times were log-transformed, because the residuals of the untransformed data of the same model do not meet the assumptions of linearity, normality (at both trial- and participant-level), and variance homogeneity (see Csibra et al., 2016 for why log-transformation is recommended for looking time data). The initial model included seven fixed-effect factors at trial-level: emotion (neutral, happy, sad, and

angry[2]), trial number (1 – 16), mean fundamental frequency (F0) (numerical variable in Hertz), intensity variation (numeral variable in dB), word duration (numerical variable in second), harmonics-to-noise ratio (HNR) (numerical variable in dB), and spectral centroid (numerical variable in Hertz). Interactions between emotion and each acoustic variable were also included. Participant-level fixed factors include sex (female = 0, male = 1) and age (numerical variable in month). To account for data dependency, the model allows random intercepts for participant, wordlist (four word orders), and first-trial-or-not (the first trial = 1, the following 15 trials = 0). Cross-level interactions of age and emotion, and sex and emotion were also included. To avoid model convergence problems, word durations and spectral centroid were rescaled. The model syntax is provided in the footnote[3].

III. Results

To achieve model parsimony, we used a deviance test to select the model with the least number of parameters (i.e., the fixed and random effect factors) that can still explain similar amounts of data variance as the initial model (Woltman et al., 2012). Both participant-level fixed-effect factors (age and sex) and their interactions with emotion were removed based on the model selection result. To demonstrate that age and sex did not explain infants' listening times to emotional speech, we ran a participant-

---

[2] This categorical variable was coded as orthogonal contrasts to avoid difficulties in interpreting interactions (i.e., emotion and acoustic variables) when treatment contrasts are used.

[3] The following syntax was used for the initial model. The de-identified data are accessible at https://doi.org/10.17605/OSF.IO/XD5AM. We dropped the main effects of participant-level factors age and sex and the cross-level interactions between emotion and age, emotion and sex in the final model (see the first paragraph in Result section).

*lmer.initial = lmer(log(Trial_LookTime) ~ 1 + Emotion + TrialNum + Emotion*f0_mean + Emotion*I(duration*1000) + Emotion*intensity_sd + Emotion*hnr_mean + Emotion*Age + Emotion*Sex + Emotion*I(spectral_centroid/10) + (1 | PID) + (1 | WordList) + (1 | Trial_1), data = data_input, REML = TRUE)*

level model with age, sex, and emotion as the only parameters and observed no significant effect. We compared and summarized the initial model, participant-level model, and the final model in Table 2. The potential effect of different word orders (four word lists) as a fixed-effect factor[4] was ruled out in a separate model. The following statistical results were from the final model fit onto log-transformed individual-trial looking times obtained from offline frame-by-frame video coding (the online individual-trial looking times yielded similar results). Paired t-tests with Bonferroni corrections were carried out to further investigate the emotion effect.

The main effects of emotion ($F(3,610) = 21.89$, $p < 0.001$), mean F0 ($F(1,622) = 81.65$, $p < 0.001$), word duration ($F(1,528) = 31.82$, $p < 0.001$), intensity variation ($F(1,625) = 4.96$, $p = 0.03$), and trial number ($F(1,593) = 41.24$, $p < 0.001$) were significant factors on infants' listening times. In general, infants' listening times were longer to the affective voices (angry, happy, and sad) than to the neutral voice (*ps* < 0.001); they listened longer to happy than angry voices ($p < .001$), and to sad than angry voices ($p = 0.003$). Infants listened more to words with lower mean fundamental frequency, to words with shorter durations (i.e., faster speaking rate), and to words with greater intensity variation. Finally, listening attention dropped as the task proceeded. Figure 1 shows the main effects of emotion, mean F0, word duration, and intensity variation. The interactions between emotion and mean F0 ($F(3,620) = 34.08$, $p < 0.001$), word duration ($F(3,624) = 11.73$, $p < 0.001$), intensity variation ($F(3,630) = 14.52$, $p < 0.001$), HNR ($F(3,630) = 38.36$, $p < 0.001$), and spectral centroid ($F(3,624) = 32.84$, $p <$

[4] $F(3,630.5)=0.35$, $p = 0.79$

0.001) were all significant. Figure 2 shows how each acoustic variable interacts with infants' listening attention to different emotions.

Table 2. F-statistics of the initial, participant-level, and final linear mixed-effect models using log-transformed looking times in individual trials of each participant as the dependent variable.

	Initial Model	**Participant-Level Model**	**Final Model**
Trial-level fixed factors			
Emotion	22.05***	2.45	21.89***
Mean Fundamental Frequency (F0)	80.44***		81.65***
Word Duration	30.88***		31.82***
Intensity variation	4.95*		4.96*
Harmonics-to-Noise Ratio (HNR)	0.55		0.36
Spectral centroid	1.63		1.32
Trial Number	41.20***		41.24***
Emotion x Mean F0	32.57***		32.94***
Emotion x Word Duration	9.47***		9.97***
Emotion x Intensity variation	9.78***		10.00***
Emotion x HNR	35.00***		35.40***
Emotion x Spectral centroid	30.24***		30.97***
Participant-level fixed factors			
Age	0.01	0.16	
Sex	0.67	1.17	
Cross-level interactions			
Age x Emotion	0.65		
Sex x Emotion	0.41		

Goodness-of-fit (Deviance)	1006.7	1428.4	1009.4

* $p < .05$, ** $p < .01$, *** $p < .001$.

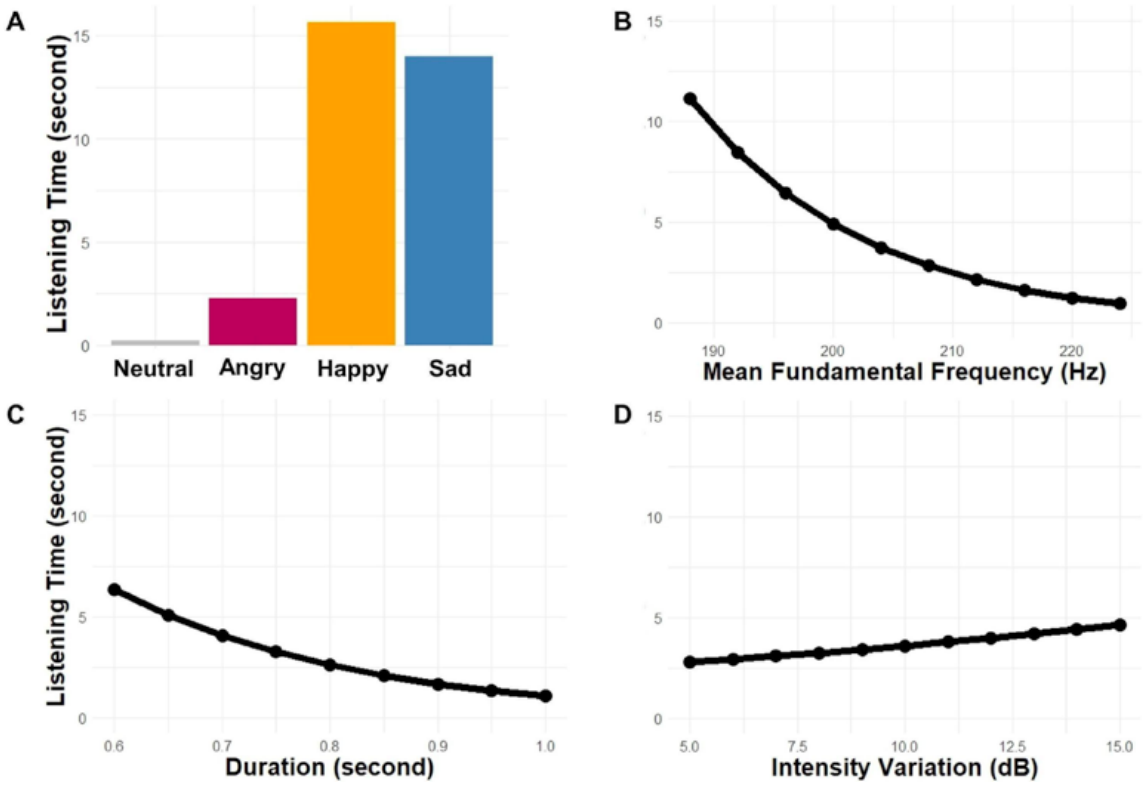


Figure 1. The model predicted listening times to different (A) emotions, (B) mean fundamental frequency, (C) word duration, and (D) intensity variation. These main effects should be cautiously interpreted because of their further interaction effects (shown in Figure 2).

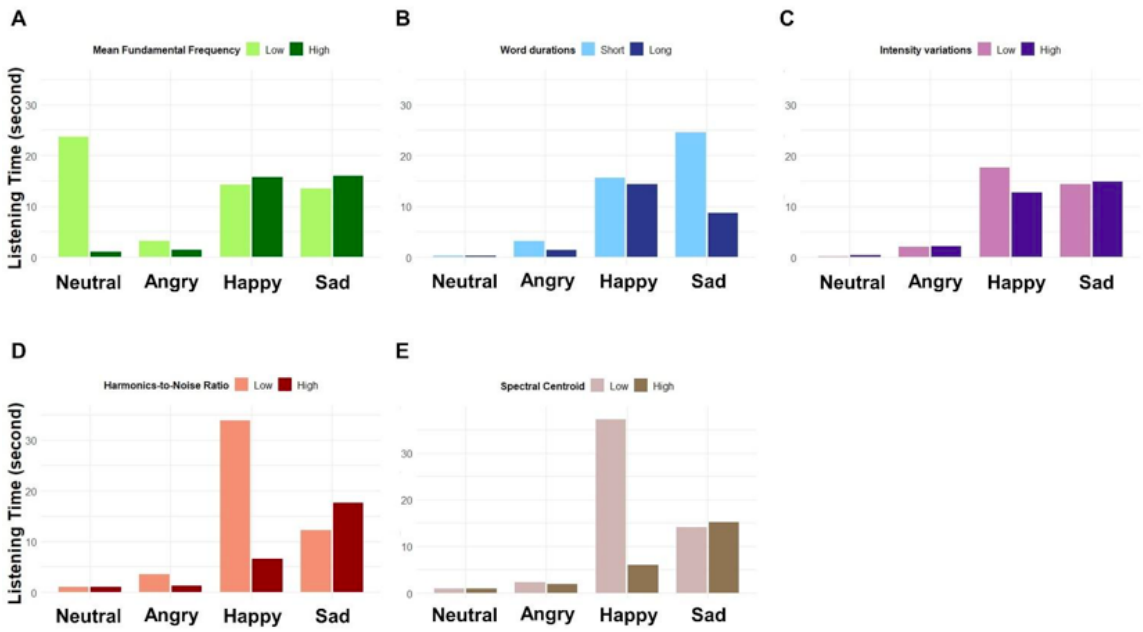


Figure 2. The model predicted listening times to different emotions modulated by (A) mean fundamental frequency (F0), (B) word duration, (C) intensity variation, (D) harmonics-to-noise ratio (HNR), and (E) spectral centroid. The 25th and 75th percentiles (from all trial-level acoustic measures) were used as high and low examples in illustrating the interactions between emotion and each acoustic variable.

The interactions between emotion and the acoustic variables were mostly observed in happy and sad prosodies (except for the mean F0 × emotion). For the interaction between mean F0 and emotion, infants only listened more to lower F0 in neutral prosody, and this is the only acoustic parameter that affected listening times to the words with neutral prosodies. In contrast, they listened longer to happy and sad sounds with a higher F0, but there was no listening difference across the two emotions. In the interaction between word duration and emotion, we observed a listening bias toward shorter words (i.e., faster speech rate) in the sad prosody but not in the other three prosodies. Sad prosody with shorter word durations even maintained longer listening attention than the happy prosody with shorter word durations. For the interaction between intensity variation and emotion, infants listened longer to happy

prosody with lower intensity variation, but not in the other three emotions. HNR indexes the amount of aperiodic signals in the speech signal. Infants listened more to the happy prosody with a lower HNR (more breathy in the speech), but they listened more to the sad prosody with a high HNR (less breathy in the speech). Sad prosody with higher HNR attracted more listening attention than happy prosody with higher HNR. Finally, there was a listening bias toward happy prosody with a lower spectral centroid, but no similar effect was observed in the other three prosodies. Similarly, sad prosody with a higher spectral centroid drew infants' attention more than happy prosody with a higher spectral centroid.

IV. Discussion

To better understand early emotional speech perception, we investigated infants' listening attention to happy, angry, sad, and neutral prosody in spoken words. Notably, we used non-repeating words to deliver the target emotions to ensure that infants responded to the emotional prosodic category, not the specific acoustic combinations of the emotion and the repeated speech stimulus. Moreover, we included five relevant acoustic variables in our analyses—mean fundamental frequency (F0), word duration, intensity variation, harmonics-to-noise ratio (HNR), and spectral centroid—to outline their roles in infants' listening attention to emotional prosody.

A. Infants Preferred the Happy Prosody

Three- to 12-month-old infants in our study showed listening preference for happy over neutral or angry prosody, which confirms previous findings indicating that infants attended more to positive affect in voices (Benders, 2013; Corbeil et al., 2013;

Singh et al., 2002). Additionally, we found that infants listened even more to happy speech with higher mean F0, less intensity variation, lower HNR, and lower spectral centroid. Higher pitch in positive affection directed infants' attention to the important social information (Soderstrom, 2007), and it is a contributor to infants' listening preference to infant-directed speech (IDS) (Fernald & Kuhl, 1987; Leibold & Werner, 2007). Because there is no comparable study on the roles of HNR, intensity variation, and spectral centroid in infants' listening attention to vocal happiness, interpretations of these observations need to be taken with caution. Breathy voices (lower HNR) may not be a common acoustic characteristic for the happy voice (Liu & Pell, 2012; Patel et al., 2011), but some breathy voice qualities were introduced by mothers during reading tasks to carry nonverbal intentions such as intimacy (Ishi et al., 2010). Less intensity variation and lower spectral centroid (i.e., less brightness in sounds) are neither common in a typical happy tone, but they may mimic the soothing voice (Fernald et al., 1984) that infants frequently hear early in life. While the current paradigm may not distinguish infants' familiarity preference from novelty preference (both manifested in longer listening times), infants tend to show more attention to the novel features in stimuli that they are exposed to more (Houston-Price & Nakai, 2004). Therefore, happy voices with these uncommon acoustic features may draw more attention because they differ from the typical happy voices that infants are familiar with. In brief, we confirmed infants' preference for happy affect in voices, and we observed infants' selective listening attention to happy voices with non-typical acoustic constituents in happy prosody.

B. Infants Did not Turn Away from the Sad Prosody

Surprisingly, infants responded similarly to both sad and happy prosodies. This result contradicts the hypothesis that they would listen less to sad than happy prosody because of their preference for the positive affect (Singh et al., 2002). Even though Singh and colleagues did not directly compare infants' listening attention between happy and sad speech, they observed longer listening times to happy than neutral sounds, and neutral than sad sounds[5], regardless of the speaking styles (IDS or ADS). One major difference was that the current study introduced another negative prosody—angry in the stimuli. The current listening task with high affective variations in roving spoken words may provide a listening context different from the context using fixed emotional pairs (Singh et al., 2002). The enriched emotional context may also encourage infants to adopt different listening strategies (Singh, 2008), especially when negative affect was included (Kiley Hamlin et al., 2010; Vaish et al., 2008).

Taking the acoustic features into account, infants' listening times to the happy and sad emotions were very close regardless of the mean F0. High or low mean F0 also did not elicit different listening patterns within happy or sad emotion, corroborating Singh and colleagues' (2002) findings on infants' similar listening times to sad speech in IDS and ADS (differed by the mean F0). Therefore, we cannot conclude that mean F0 plays a major role in driving infants' differential attention to sad and happy prosody. Neither did this result support our second exploratory hypothesis that infants would listen more to sad sounds because it shares longer word durations (slower speech rate)

[5] Except that infants listened similarly to neutral ADS and sad IDS without showing a preference for a relatively positive affect.

with the IDS (Fernald & Simon, 1984; Fernald et al., 1989; Stern et al., 1982). Instead, infants only listened more to sad speech when the word durations were short, indicating that faster speech rate could better maintain infants' listening attention to sad prosody. This effect of duration in the sad speech was not observed in Singh and colleagues' report (2002) when ADS and IDS were compared. Because the current report only used sad ADS, it is possible that the attention-maintaining role of a faster speaking rate can only be observed in this listening context. Except for shorter word durations, sad speech with a higher HNR (less breathy) and spectral centroid (brighter sound) attracted infants' attention more than the happy speech with similar HNR and spectral centroid measures. Higher HNR and spectral centroid are two acoustic characters (out of many) of happy sounds. While it may be an over-statement to conclude that brighter voices with less breathy quality introduce some positive affect into the sad speech, perhaps both acoustic characters make the sad ADS less sad-sounding and more intriguing to infants. Although there is a lack of similar empirical studies for a direct comparison, our report on infants' listening attention to sad sounds and the modulating roles of word durations, HNR, and spectral centroid provided some evidence for future studies to test directly.

C. Infants Listened Less to Angry Prosody Irrespective of the Acoustic Features

We did not observe an age effect in infants' responses toward happy and angry prosodies as predicted. Instead, all infants paid more attention to the happy than angry prosody. Given that the two vocal emotions share similar acoustic features and were presented over non-repeating words, it is surprising that three- to 12-month-old infants in the current study could still respond to the two differently. In the study by Mastropieri

and Turkewitz (1999), who presented angry and happy speech from four female speakers to newborns, the newborns were able to generalize across speakers and form two emotional prosodic categories. Taking the Mastropierir and Turkewitz (1999) together with ours, we believe that infants before the age of one can extract emotional prosodic categories over various non-repeating examples and differentiate between happy and angry voices, and they show the listening preference for happy prosody right after birth. The five acoustic variables did not modulate infants' listening times to angry prosody, indicating that infants' lack of interest in angry sounds could not be recovered by any of the included acoustic features. This less attention to high-arousal negative speech was in line with the study showing infants' looking preference for happiness to anger when audiovisual emotional information was presented (Soken & Pick, 1999).

D. Neutral Tone Was the Least Interesting Prosody Unless It Is with a Lower F0

Neutral prosody attracted the least listening attention compared with the other three emotional prosodies, except when delivered at a lower F0. We initially included neutral prosody as a reference, so we did not expect to see any effects of the acoustic variables. Infants' preference for neutral speech with a lower F0 also seemed to conflict with the literature showing infants' preference for a higher F0 (Fernald & Kuhl, 1987; Masapollo et al., 2016; Trainor & Zacharias, 1998). However, the literature on infants' listening bias to a higher F0 was usually conducted in IDS, different from the context of adult-directed speech we used in the current study. Moreover, our high F0 example was around 223 Hz, which was used as a low F0 example in the previous study (Trainor & Zacharias, 1998). It is likely that our low F0 example (188.4 Hz) was not tested in

previous infant preferential listening studies. To sum up, infants' short listening time to the neutral prosody rather than the affective prosody was expected, as socio-emotional information is crucial in early language environments (Kuhl, 2007). The role of mean F0 in infants' neutral speech perception will need future research to elaborate and clarify.

E. No Age or Sex Effect in Early Emotional Speech Processing

The lack of an age effect in infants' vocal emotion processing for the four emotional prosodies suggests that younger infants in the current study demonstrated similar listening patterns as the older infants. Our finding here was not in line with previous reports (e.g., Flom & Bahrick, 2007), and this divergence is likely related to different testing protocols and speech stimuli. Previous studies demonstrated an increased auditory sensitivity to emotional voices with age using the habituation paradigm, in which infants were familiarized with one vocal emotion and tested on a new emotional category to see if they can detect the change of switching from one category to the other (Flom & Bahrick, 2007; Walker-Andrews & Grolnick, 1983; Walker-Andrews & Lennon, 1991). To measure infants' change detection response, the acoustic differences between the familiarized and tested emotional speech must surpass infants' internal discriminatory criteria. Under this condition, younger infants would not show emotional prosody change detection if they cannot differentiate between the specific emotional contrast carried by the repeated lexical content (e.g., angry and happy are both high-arousal and hard to be differentiated). Our experimental design did not use the habituation paradigm to test simple discrimination; instead, we included non-repeating lexical items in each emotional prosody that would tap into perceptual

abstraction/grouping across multiple entries to establish and compare the four different vocal emotional categories. The use of four vocal emotions in a single central-fixation task rather than two emotions in a standard habituation task was intended to encourage young infants to form different emotional categories based on subtle acoustic differences (e.g., happy and angry). The affective cues may facilitate young infants' attention to similar emotional voices that may be missed in a change-detection task.

We additionally examined the effect of biological sex in early emotional prosody speech perception, but no significant effect was found. This result is not surprising because neither did a previous vocal emotional discrimination study observe a sex effect in infants (Walker-Andrews & Grolnick, 1983). If their relatively simple emotional sound discrimination task did not reveal a sex effect, it might be unexpected to see a sex effect in our more complex experiment with four emotional voices. Even though one report observed mothers using different prosodic features in their speech to male and female infants (Kitamura & Burnham, 2003), our data suggested that differences in prosodic inputs may be unidirectional from the caregivers rather than contingent on infants' distinct responses. While later studies observed sex differences in emotional prosody processing in early adolescence (Fujisawa & Shinohara, 2011) and adulthood (Schirmer et al., 2002), it is possible that these differences emerge with repeated exposure to qualitatively distinct socio-emotional inputs. Together, we propose that the different emotional processing across males and females may be a product of very large or long-term differences in the learning environments.

V. Limitations and Future Directions

There are some limitations to the current study. First, the age range of the infants was broad, so the current sample size may be relatively small to well represent infants of different developmental stages before the age of one. In order to capture the potential age effect, future work should either focus on a narrower age range or carefully recruit more infants in each age group to better characterize the processing differences across infancy. For instance, five- and seven-month-old infants started to match audiovisual emotions (Soken & Pick, 1992; Walker-Andrews, 1986, 2008), indicating an emotional appraisal that is more advanced than emotional perception. Targeting these two age groups and recruiting more participants in each group may provide a more fine-grained view of the developmental trajectory of emotional speech processing. Second, emotional prosody is a complex signal characterized by more than the five acoustic parameters as analyzed and reported in our study. Further investigations are needed to establish the optimal models in search for the acoustic correlates for infants' preferential behaviors of emotional speech perception. Third, we used emotional adult-directed speech (ADS), not the commonly used infants-directed speech (IDS), to measure infants' selective attention to emotional voices. From the stimulus end, the acoustic profiles of the same emotion are similar across ADS and IDS (Trainor et al., 2000). From infant listeners' end, their listening times to the same emotion in ADS and IDS are similar (Singh et al., 2002). Therefore, we may expect similar, if not more distinct, effects of emotion and acoustic variables on infants' listening attention when IDS is used. Follow-up studies using emotional IDS over phonetically balanced words can provide empirical evidence to

strengthen the notion that vocal affect and its functions are relatively independent of the speaking style.

The current study fits into a bigger picture of the interplay between socio-emotional and language development in infancy and childhood, especially in populations such as children with autism spectrum disorder (ASD), developmental language disorder (DLD), and cochlear implants (CIs). Children with ASD may tell the acoustic differences across vocal emotions, but they generally struggle with emotional voice appraisal (McCann & Peppé, 2003; Zhang et al., 2021). They also show less orientation to sounds with social information and may therefore miss the enriched speech inputs for language learning (O'connor, 2012). Children with DLD also struggle with emotion processing, and a recent study was supportive of the idea that socio-affective processing skills and language skills mutually affect one another in this population (Bahn et al., 2021). Cochlear implants provide invaluable early auditory inputs for children with congenital hearing loss, but the implants deliver degraded spectral information—the crucial acoustic features of both linguistic and emotional prosody (Jiam et al., 2017). Therefore, understanding young listeners' attention to emotional speech and the consequential effect on language learning may elucidate the atypical language development in children with CIs. To this day, the connections between socio-emotional and language development are still far from clear. Future studies on speech perception and language learning can be designed to include natural emotional prosody contrasts in the speech materials for investigating how socio-emotional speech input may shape language development in these special populations.

VI. Conclusion

In summary, typically developing infants at 3~12 months of age showed distinct patterns for happy, sad, angry, and neutral prosodies in spoken words with a generally longer listening time for happy and sad prosodies, and the least interest in the neutral prosody. Furthermore, mean F0, word duration, intensity variation, HNR, and spectral centroid each played a significant role in infants' listening attention to emotional voices, which varies depending on the emotion category. With our block stimulus design of roving spoken words, no age or sex effects were observed. These results provide direct evidence for the influences of four vocal emotion categories and five acoustic parameters on infants' listening attention for emotional speech in the first year of life, which have implications for further studies on socio-affective development and language learning in typically developing children as well as children with problems in emotional prosody perception.

Chapter 3: Infants' Neural Sensitivity to Emotional Prosody Differences in Spoken Words (Study 2)

I. Introduction

Emotional prosody (i.e., vocal emotion) plays an essential role in infants' first year of life. Affection in the human voice provides social connections and helps regulate infants' behaviors (Grossmann, 2010). The regulating role of emotional prosody is crucial toward the end of infants' first year of life when they start modifying behaviors based on caregivers' vocal emotions accordingly (Mumme et al., 1996; Vaish & Striano, 2004). Emotional prosody also attunes infants' attention to relevant speech inputs and is essential in early language development under the socio-emotional framework (Hohenberger, 2011). In this framework, decoding the affective prosody information in speech is the first step toward language learning before infants can use the language-specific linguistic information. In other words, infants before the age of one rely more on the paralinguistic cues than the linguistic ones (Fernald, 1989, 1993; Lawrence & Fernald, 1993). Therefore, timely processing of emotional speech prosody is critical in guiding infants' behaviors, social interactions, and later language learning.

To understand infants' timely processing of emotional prosody in speech, scientists have used electroencephalography (EEG) to measure their neurophysiological responses (Cheour et al., 2000; Csibra et al., 2008; de Haan, 2002). EEG records listeners' neural activities during cognitive tasks at the millisecond scale, and the averaged neural responses time-locked to the task events (e.g., sound stimuli), known as event-related potentials (ERPs), can be analyzed to assess listeners' sensory and cognitive processing of the presented stimuli even without attentional focus on auditory

inputs. In fact, ERP has been employed in numerous studies to reveal infants' differential responses to various sounds without their overt behavioral reactions, making it a convenient and temporally precise tool in the early processing of emotional prosody. However, the neural mechanisms underlying infants' vocal emotion decoding are far from clear. Developmental research on emotional information processing has been mainly focused on facial rather than vocal expressions of emotions (Grossmann, 2010; Morningstar et al., 2018). For instance, very little is known about how infants extract different emotional prosody categories in spoken words, which would presumably show age-dependent and category-specific changes similar to their visual responsiveness to facial expressions of emotion (Leppänen, Moulson, Vogel-Farley, & Nelson, 2007; Nelson & De Haan, 1996). The current study aims to fill this gap by measuring infants' pre-attentive neural responses to various basic vocal emotions. Successful elicitation and analysis of distinct ERPs for different emotional prosodies would add empirical evidence to deepen our understanding of the neural correlates underpinning vocal emotional processing in infancy.

Infants' Emotional Prosody Discrimination through Behavioral Research

Infants' behavioral sensitivities to different emotional voices emerge early, but they may not reliably show different behavioral reactions to distinct voices even until seven months of age (Soken & Pick, 1992, 1999; Walker-Andrews, 1986). Newborns are already sensitive to prosodic information in speech (Mehler et al., 1978; Moon et al., 1993). For instance, one early study revealed that newborns opened their eyes more to the happy voice, but not angry, sad, or neutral voices when listening to their native

language (Mastropieri & Turkewitz, 1999). This study indicates that newborns already show some degrees of vocal emotion discrimination ability. Another study showed that three-month-old infants could detect an emotional sound change from sad to happy by resuming their attention to the changed sound presentation, but not when the sound changed from happy to sad (Walker-Andrews & Grolnick, 1983). This order-specific behavioral response confirms that three-month-old infants have emotional voice discrimination ability (at least between happy and sad voices), but their behavioral responses cannot reflect this voice sensitivity consistently. On the other hand, five-month-old infants can reliably differentiate between more emotional prosodies (including anger) regardless of the sound presentation order, but they can only do so when an image of a human face is displayed along with the sound (D'Entremont & Muir, 1999; Walker-Andrews & Lennon, 1991). In other words, five-month-old infants' sensitivities to different emotional prosodies can only be reliably detected with a relevant visual stimulus, not auditory inputs alone. Along the developmental timeline, seven-month-old infants start to show more reliable vocal emotion discrimination, but their emotional voice sensitivities are still susceptible to visual distractors such as upside-down faces (Soken & Pick, 1992, 1999; Walker-Andrews & Grolnick, 1983; Walker, 1982). Taken together, while newborns are already sensitive to different emotional prosodies, infants' overt behavioral responses to different vocal expressions may not be reliably elicited even when they are seven months of age.

The behavioral assessment of infants' voice discrimination is usually carried out in the habituation paradigm (Colombo & Mitchell, 2009; Fantz, 1964; Groves &

Thompson, 1970). For testing vocal emotion discrimination, researchers would first present one emotional voice repetitively to the infant listeners. Once the infants are habituated by the emotional voice and show decreased interest, researchers would play a new vocal emotion and observe if infants resume their attention to the new sounds. To successfully pass the habituation task, infants first register the emotional voice differences in their auditory system and then demonstrate that they hear the differences by showing distinct behavioral reactions such as voluntary eye fixation (Aslin, 2007). Unfortunately, the habituation paradigm can only tell if infants behaviorally show distinct responses to the sounds. This behavioral paradigm cannot distinguish infants whose auditory system does or does not register the sound differences if both groups show similar behavioral responses to the habituated and new emotional categories. In other words, a failure to show increased attention to the new vocal emotional category does not necessarily mean that infants cannot tell the two vocal expressions. Instead, perceptual-irrelevant factors (e.g., inherent preference for particular stimuli) may affect infants' behavioral reactions to different voices (Oakes, 2010). Therefore, a more sensitive measurement is needed to capture infants' early emotional voice discrimination ability when infants show inconsistent behavioral responses.

EEG recordings have proven to be a sensitive tool in developmental cognitive science that complements behavioral measurements (De Haan, 2007; Hartkopf et al., 2019). Since infants' neurophysiological responses can be measured without their behavioral reactions, ERPs to vocal emotions may be able to reveal the differences in infants who process the sounds differently but score similarly in behavioral tasks. By

including infants' neurophysiological responses to emotional prosodies, we can further address young infants' unreliable behavioral discrimination of emotional speech and better understand the age-dependent changes in decoding the crucial affective information that influences their socio-emotional and cognitive development (Hohenberger, 2011).

Developmental ERPs to Emotional Prosody in Infancy

Neurophysiological measurements have been broadly used to understand developmental auditory processing (Cheour, 2007; Csibra et al., 2008). Among the relevant auditory ERPs, the mismatch response (MMR) is the neural marker that indexes infants' pre-attentive voice discrimination (de Haan, 2002; Kushnerenko et al., 2013). MMR is usually elicited by the infrequent sound (*Deviant*) in a continuous stream of the same frequent sound (*Standard*) (Garrido et al., 2009; Näätänen et al., 2007). MMR appears as a negative deflection around 150 to 250 ms after sound onset in adult listeners (so-called mismatch negativity, MMN). However, MMR's time window and polarity in infants vary widely depending on the stimulus type and the infant's age (Cheour et al., 2002; Csibra et al., 2008; Friederici et al., 2002; Maurer et al., 2003). Unlike adults' MMN, infants' MMR mostly appears as a slow positive wave at a later window around 200 – 450 ms post-stimulus (e.g., Cheng et al., 2012; Dehaene-Lambertz, 2000; Friederici et al., 2002; He et al., 2009; Leppänen et al., 2004; Winkler et al., 2003). Many infant EEG studies have already successfully recorded infants' MMR to different speech and voices (Cheour-Luhtanen et al., 1995; García-Sierra et al., 2021; Shafer et al., 2012; Wanrooij et al., 2014). Along with the low requirement of listeners' active

engagement in the task, the MMR paradigm is suitable for investigating infants' neural sensitivities to emotional prosodies in speech.

The literature on early neurophysiological responses to emotional prosody is quite limited (Kok et al., 2014; Morningstar et al., 2018), but a few neuroimaging reports suggested that infants before the age of one already show different neural activities to voices with different emotions (Blasi et al., 2011; Grossmann et al., 2005; Minagawa-Kawai et al., 2011). Studies focusing on MMR to emotional prosodies even observed newborns' differential neural sensitivities to emotional voices presented over simple syllables (Cheng et al., 2012; Kostilainen et al., 2020; Kostilainen et al., 2018; Zhang et al., 2014). For instance, neonates' MMRs to angry, happy, and fearful sounds differ around 300 to 500 ms after the sound onset (Cheng et al., 2012; Zhang et al., 2014), with stronger MMR to the happy than the angry sounds (Kostilainen et al., 2020). Cheng and colleagues (2012) further demonstrated that newborns' emotional MMRs are not merely driven by the acoustic differences by showing no clear MMRs to acoustically-matched non-speech emotional stimuli.

While infants' neural sensitivities to different emotional prosodies are present at birth, the developmental trajectory of this neural mechanism remains unclear as there have been very few reports on older infants' MMRs for auditory processing of emotion. The closest report was from Grossmann et al. (2005), demonstrating that seven-month-old infants show ERPs to happy and angry voices. One limitation is that previous auditory emotional MMR tasks all used simple and fixed syllables to carry the vocal emotional expression. Whether or not infants' pre-attentive system can discriminate

between different emotional voices and automatically extract emotional categories based on statistical regularities (such as fundamental frequency patterns associated with different vocal emotions) in a more complex linguistic context (e.g., spoken words) has not been investigated. Another less-considered factor is whether sex differences present in infants' neural responses to the emotional prosody, which may also show age-dependent and emotional-category-specific effects. Previous newborn studies did not find a sex effect on emotional MMR (Cheng et al., 2012; Kostilainen et al., 2018), but adult studies have (Hung & Cheng, 2014; Schirmer et al., 2005). Thus, it would be of great value to address the developmental trajectory of the sex effect on infants' pre-attentive neural processing of emotional prosody in natural speech.

The Current Study

The current study aims to investigate whether infants before the age of one year can automatically extract different emotional prosodies in non-repeating spoken words and how this ability develops as a function of age and sex. To this end, we employed a roving multi-feature oddball paradigm, which can record and compare listeners' neural sensitivities to multiple types of emotional sounds (i.e., happy, angry, and sad) in a single session. The multi-feature oddball task may be challenging for infants because of the increased number of novel auditory events for their auditory system to detect. However, previous reports have successfully observed newborns' MMRs to three emotional *Deviants* along with six other acoustic *Deviants* (Kostilainen et al., 2020; Kostilainen et al., 2018).

The most significant modification we applied to the current emotional multi-feature oddball task was using non-repeating spoken words to deliver the emotional prosody. The rationale was to examine whether infants' pre-attentive neural system can detect the overarching emotional prosodic category over the varying linguistic contents based on statistical regularities of acoustic cues for distinct vocal emotion categories. Due to the high acoustic variations within the same vocal emotion in this setup, we expect to see more subtle emotional MMRs in the early (100 – 200 ms; e.g., Kostilainen et al., 2018) and late windows (300 – 500 ms, e.g., Cheng et al., 2012; Zhang et al., 2014) after the sound onset. Even though Kostilainen and colleagues (2018) did not observe distinct early MMRs to different emotional voices in newborns, we expected that our infant listeners in the broader and older age range of 3-11 months would be able to show different early emotional MMRs due to auditory experience in their learning environment. In addition, we expected to find category-specific differences such as stronger late MMRs to anger than other emotional prosodies as previous reports suggest that the neural system automatically orients to threat-related signals (Cheng et al., 2012; Grossmann et al., 2005; Grossmann et al., 2006). In this vein, angry and happy voices may elicit distinct MMRs (Grossman et al., 2005), even though infants may not distinguish these two vocal emotions reliably in behavioral tasks (Flom & Bahrick, 2007; Walker-Andrews & Lennon, 1991). Although infants' auditory MMRs to sadness have not been systematically compared with other vocal emotions in the literature, we predicted that their sad MMRs would be different from happy and angry MMRs because three-month-old infants can already differentiate sad voices from other emotions in

behavioral tasks (Walker-Andrews & Grolnick, 1983). Finally, sex-specific differences in the auditory emotional MMRs may emerge. It is possible that more adult-like MMR (i.e., MMN) may show up in female infants than male infants because female infants already show better visual emotional processing (facial expressions, McClure, 2000). However, due to the lack of systematic investigation of sex differences in vocal emotional processing in infancy, the search for potential sex effects on vocal emotion processing remained exploratory in the current study.

II. Method

A. Participants

The final sample included 42 infants between the ages of two months 26 days and 11 months 11 days (male = 22, female = 20; mean age = 7.5 months or 228 days). Forty-six typically developing infants from 3 to 12 months of age (male = 25, female = 21; mean age = 7.6 months or 229 days) were recruited through advertisements, words of mouth, and the infant participant pool of the Institute of Child Development at the University of Minnesota. All infants were born full-term (38 – 42 weeks), healthy with normal hearing, and from English-speaking families. Four infants' data were not included due to the EEG cap being pulled off (n = 1), crying (n = 1), and equipment failure (n = 1). The experimental protocol was approved by the local Institutional Review Board. Parents signed the informed consent for their children prior to the participation and received $20 as monetary compensation upon completion.

B. Stimuli

All speech stimuli were taken from the Toronto Emotional Speech Set (Dupuis & Pichora-Fuller, 2010), including 200 monosyllabic phonetically balanced words (Northwestern University Auditory Test No. 6, NU-6; Tillman & Carhart, 1966) as listed in Appendix B. Each of the 200 words was spoken in neutral, happy, sad, and angry voices by a young female speaker, yielding a total of 800 stimuli. The sounds were sampled at 24414 Hz, with the mean sound intensity levels equalized using Praat 6.0.40 (Boersma & Weenink, 2020). Table 3 summarizes the mean fundamental frequency (F0), duration, intensity variation, harmonics-to-noise ratio (HNR), and spectral centroid in each emotional prosody. These five acoustic measures are commonly used to characterize different vocal emotions (Amorim et al., 2019; Banse & Scherer, 1996; Johnstone & Scherer, 2000; Mani & Pätzold, 2016).

Table 3. **The acoustic properties of each emotional prosody**

Emotions	*Mean F0 (Hz)*	*Duration (ms)*	*Intensity Variation (dB)*	*HNR (dB)*	*Spectral centroid (Hz)*
Angry	216.71	646	11.15	9.22	1810.96
Happy	226.13	742	10.82	17.53	1052.92
Sad	180.42	822	10.18	19.31	408.79
Neutral	195.04	667	9.14	18.75	758.43

Note. The averaged values of all the words were used to report the mean fundamental frequency (F0), word duration, intensity variation, harmonics-to-noise ratio (HNR), and spectral centroid in each emotional prosody.

C. Procedure

We adopted a multi-feature oddball paradigm (or optimal paradigm; Näätänen et al., 1978; Näätänen et al., 2004; Thönnessen et al., 2010) to examine infants' early neural sensitivities to happy, angry, and sad prosodies as against the neutral prosody. The multi-feature oddball paradigm is a passive listening protocol that allows us to measure three emotional prosody contrasts (from neutral to angry, from neutral to happy, and from neutral to sad) within the same recording session, which is suitable for infant participants. We presented 600 trials in total. The *Standard* stimuli were words in a neutral tone (presented with 50% probability, 300 trials). The three emotional tones (happy, angry, sad) served as three *Deviant* stimuli (each presented with 16.7% probability, 100 trials). For the 300 *Standard* trials, all the 200 words in neutral voice were used, and 100 words were randomly selected for repetition. For each type of the 100 *Deviant* trials, 100 words in each emotional voice were randomly selected and used. The sounds were always presented in alternating *Standard* and *Deviant* fashion. The three types of *Deviant* (three emotions) were pseudo-randomly interspersed, with no consecutive *Deviant* trials in the same emotional prosody (see Figure 3 for an example of the sound presentation order). The inter-stimulus interval (ISI) was randomized between 800 – 900 ms, and the total recording time was around 25 minutes.

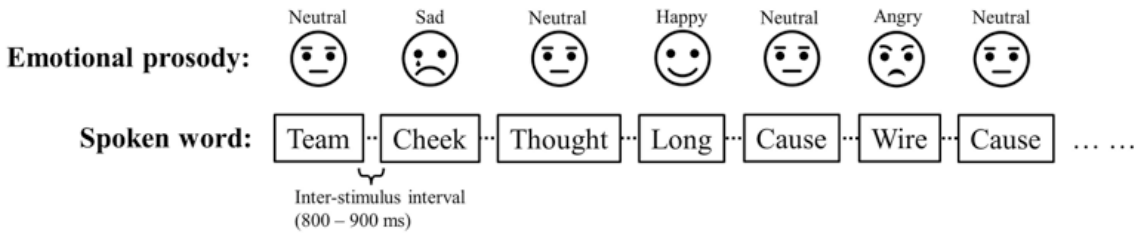


Figure 3. A schematic example of the order of the trials. The *Standard* (neutral prosody) and *Deviant* (angry, happy, and sad prosodies) were always alternating, and the three emotions (*Deviants*) were pseudo-randomly interspersed.

Infants were seated in their parents' laps in an electrically and acoustically treated booth (ETS-Lindgren Acoustic Systems) with a 64-channel WaveGuard EEG cap. One research assistant stayed in the booth and played with silent toys to entertain the infants. A television displaying silent cartoons was also on to keep the infants engaged and still. Parents were instructed to ignore the speech sounds and soothe their children during the EEG recording session. The speech sounds were played via two loudspeakers (M-audio BX8a) placed at a 45-degree azimuth angle three feet away from the participants and presented at 65 dB SPL at the subject's head position (Zhang et al., 2011), which was calibrated prior to the experiment using a standard 1000 Hz tone. The sound presentation was controlled by E-Prime (Psychological Software Tools, Inc) using a Dell PC outside the sound-treated room. Continuous EEG data were recorded through the Advanced Neuro Technology EEG System (Advanced Source Analysis version 4.7). The WaveGuard EEG cap has a layout of 64 Ag/AgCl electrodes following the standard International 10-20 Montage system with intermediate locations, and it is connected to a REFA-72 amplifier (TMS International BV). The default bandpass filter for raw data recording was set between 0.016 Hz to 200 Hz, and the sampling rate was 512 Hz. The electrode AFz served as the ground electrode. The impedance of all electrodes was kept under 10 kΩ.

D. Data analysis

The continuous EEG data preprocessing was complete offline by EEGLAB v14.1.1 (Delorme & Makeig, 2004). The continuous EEG data were low-pass filtered at 40 Hz, downsampled to 250 Hz, and high-pass filtered at 0.5 Hz. The EEG data were then re-referenced to the average of the two mastoid electrodes. Next, we applied the "Clean Rawdata" EEGLAB plug-in to help remove low-frequency drifts and non-brain activities (e.g., muscle activity, sensor motion). Data were then decomposed by the Independent Component Analysis (ICA) algorithm (Dammers et al., 2008; Delorme et al., 2001) to attenuate influences from eye blinks and other artifacts. ERP epochs were extracted from 100 ms pre-stimulus onset to 1000 ms post-stimulus onset, and baseline correction was applied using the mean voltages of the 100-ms baseline period. Epochs containing data points over the range of 150.0 μV were rejected before averaging. Using ERPLAB v7.0.0 (Lopez-Calderon & Luck, 2014), event-related potentials (ERPs) were derived for *Standard* (neutral prosody) and each three types of the *Deviant* (angry, happy, and sad prosodies). Difference waveforms were created by subtracting the *Standard* ERP from each *Deviant* ERP, yielding happy, angry, and sad difference waveforms. The data from infants with fewer than 30 trials in any of the *Standard* or *Deviants* conditions were removed from further analysis (male = 5, female = 5; mean age = 7.1 months or 218 days).

All statistical analyses were completed in R (https://www.r-project.org/) with the packages "lme4" (Bates et al., 2015), "lmerTest" (Kuznetsova et al., 2017), and "emmeans" (Lenth et al., 2018). The difference waveforms were used for assessing two

target components—early mismatch response (early MMR, 100 – 200 ms post-stimulus onsets) and late mismatch response (late MMR, 300 – 500 ms post-stimulus onsets). The time window for each MMR was selected based on previous EEG studies on emotional prosody perception in newborns and infants (Grossmann et al., 2005; Kostilainen et al., 2020; Kostilainen et al., 2018; Zhang et al., 2014) and visual inspection of the current grand average difference waveforms. Since clear local peaks may not be present in infants' difference waveforms, the early and late MMR amplitudes for later statistical analyses were calculated using the mean voltages of the whole target windows (early MMN, 100 – 200 ms; late MMR, 300 – 500 ms). The amplitudes were calculated for channels at frontal (F-line, F3, Fz, F4), central (C-line, C3, Cz, C4), and parietal (P-line, P3, Pz, P4) regions (Kostilainen et al., 2018). These amplitudes were then used as the dependent variables in the later statistical models.

Two linear mixed-effect models were respectively performed on early and late MMR amplitudes. Each model included a by-participant intercept as a random-effect factor. Deviant/Emotion (happy, sad, and angry), region of the electrode (anterior, central, and parietal), and laterality of the electrode (left, middle, and right) were included as trial-level fixed-effect factors. Infants' biological sex (female and male) and age (in month, numerical) were included as participant-level fixed factors. Finally, cross-level interactions of emotion and sex and emotion and age were also included. Post-hoc t-tests with Bonferroni corrections ($\alpha = .05$) were conducted to characterize any significant interaction.

III. Results

The linear mixed-effect models were fitted onto the average amplitudes of early MMR and late MMR extracted from the difference waveforms, which were derived from subtracting the *Standard* ERP from the *Deviant* ERPs. To demonstrate the developmental trends of the neural sensitivities to emotional prosody, the grand mean ERP waveforms and difference waveforms of younger and older infants (using median-split) recorded from frontal (F-line, F3, Fz, F4), central (C-line, C3, Cz, C4), and parietal (P-line, P3, Pz, P4) electrodes for all emotional prosodies are displayed in Figures 4 and Figure 5. The topographic maps of each emotional prosody's early and late MMRs are presented in Figure 6.

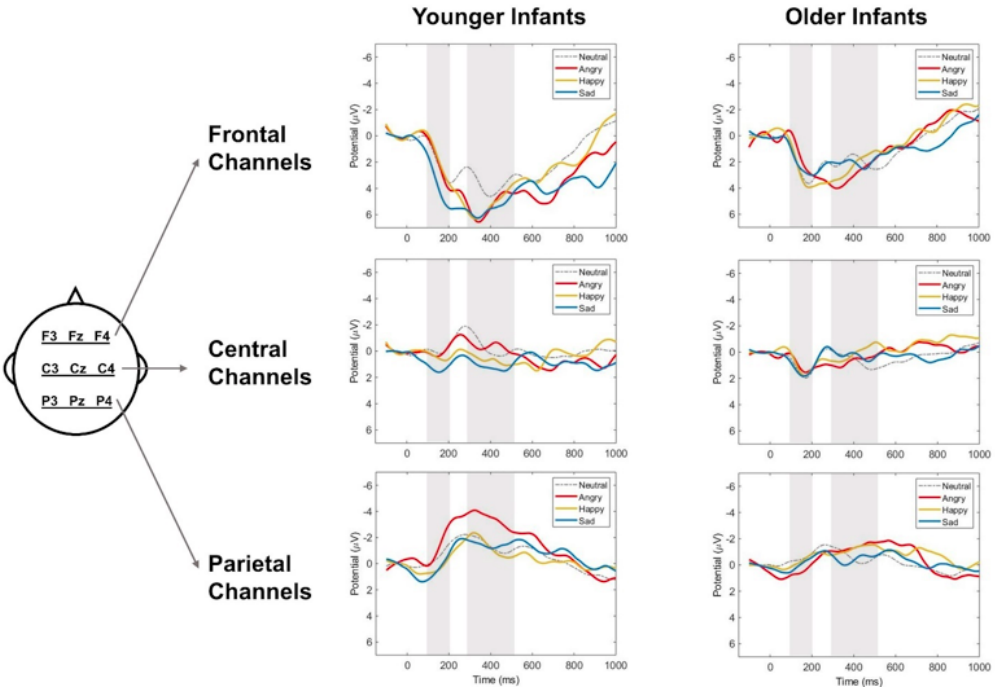


Figure 4. The grand mean event-related potential (ERP) waveforms of *Standard* (neutral prosody) and *Deviants* (angry, happy, and sad) in younger and older infant listeners (split by median age at 8.2-month). Mean amplitudes of the F-line (F3, Fz, F4), C-line (C3, Cz, C4), P-line (P3, Pz, P4) electrodes were used for the waveforms. The gray shaded areas mark the windows for early mismatch response (early MMR, 100 – 200 ms) and late MMR (300 – 500 ms).

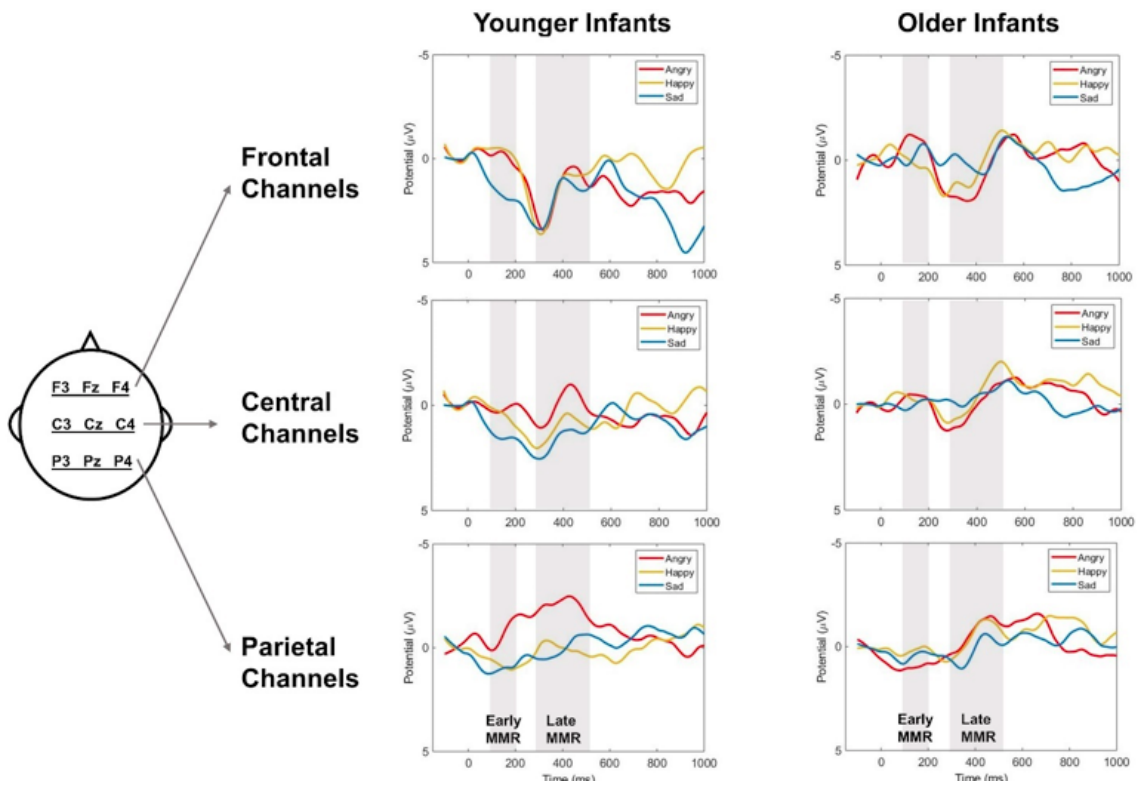


Figure 5. The grand mean difference waveforms (*Standard* waveforms subtracted from *Deviant* waveforms) of angry, happy, and sad in younger and older infant listeners (split by median age at 8.2-month). Mean amplitudes of the F-line (F3, Fz, F4), C-line (C3, Cz, C4), P-line (P3, Pz, P4) electrodes were used for the waveforms. The gray shaded areas mark the windows for early mismatch response (early MMR, 100 – 200 ms) and late MMR (300 – 500 ms).

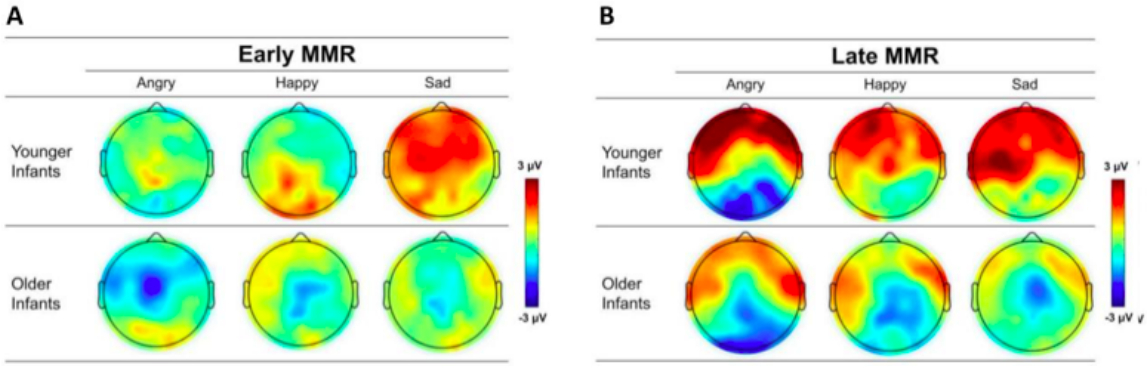


Figure 6. The scalp topographic maps of (A) early mismatch response (MMR) and (B) late MMR to angry, happy, and sad emotional prosodies averaged across younger and older infant listeners (split by median age at 8.2-month). The topographies are based on the average values in each component window (early MMR, 100 – 200 ms; late MMR, 300 – 500 ms).

A. Early Mismatch Response (Early MMR)

The main effect of emotion ($F(2,818) = 14.75$, $p < 0.001$) was significant, but not the main effects of electrode region ($F(2,817) = 1.05$, $p = 0.35$) or electrode laterality ($F(2,817) = 1.78$, $p = 0.17$). In general, early MMR amplitudes to sad prosody were more positive than happy ($p = 0.03$) and angry prosodies ($p = 0.002$). The participant-level main effect of sex was not significant ($F(1,32) = 2.11$, $p = 0.16$), neither was the main effect of age ($F(1,32) = 0.74$, $p = 0.39$). However, there was a significant interaction between emotion and age ($F(2,818) = 11.42$, $p < 0.001$), with early MMR values to happy and sad prosodies going more negative with age and early MMR values to angry going more positive with age. The model is summarized in Table 4, and Figure 7 shows the interaction effect of emotion and age on infants' early MMR amplitudes.

Table 4. Summary of the linear mixed-effect model using the amplitudes of early MMR as the dependent variable.

Factor	*Numerator df*	*Denominator df*	*F*	*p*
Trial-level fixed factors				
Emotion	2	818	14.75	< .001 ***
Region	2	817	1.05	0.35
Laterality	2	817	1.78	0.17
Participant-level fixed factor				
Age	1	32	0.74	0.39
Sex	1	32	2.11	0.16
Cross-level interaction				

Emotion * Age	2	818	11.42	< .001 ***
Emotion * Sex	2	817	2.64	0.07

* $p < .05$, ** $p < .01$, *** $p < .001$.

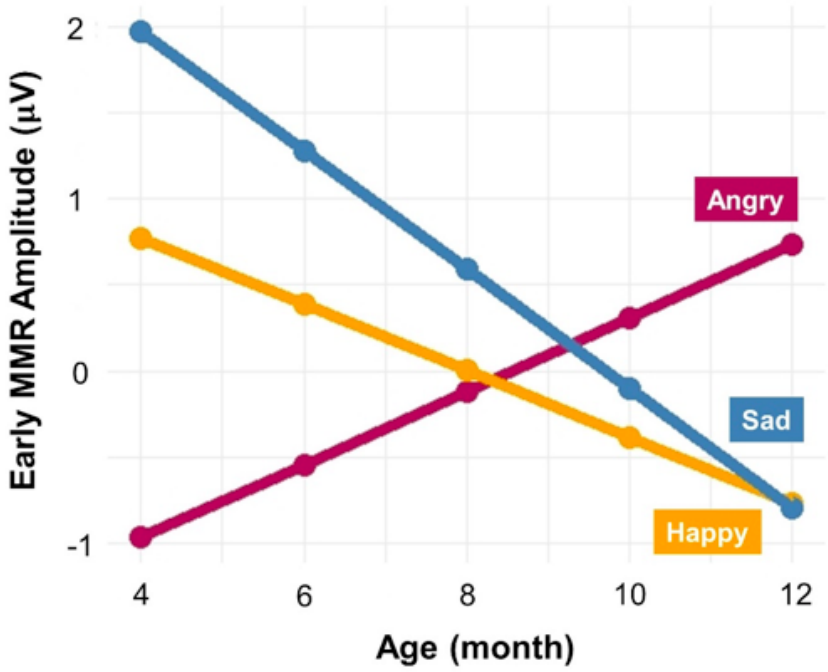


Figure 7. The interaction effect of emotion and age displayed in the model predicted MMN amplitudes to angry, happy, and sad emotional prosodies.

B. Late Mismatch Response (Late MMR)

The main effects of emotion ($F(2,800) = 12.12$, $p < 0.001$) and electrode region ($F(2,800) = 19$, $p < 0.001$) were significant, but not the main effect of electrode laterality ($F(2,799) = 2.73$, $p = 0.07$). In general, late MMR to the angry prosody was more negative than the sad prosody, but the post-hoc t-test only approached significance level after Bonferroni correction ($p = 0.08$); and frontal channels and central channels recorded more positive late MMR than the parietal channels ($ps < 0.001$). The

participant-level main effect of sex was significant ($F(1,32) = 5.45$, $p = 0.03$), with male infants showing more positive late MMR. The main effect of age was not significant ($F(1,32) = 0.61$, $p = 0.44$), but there was a significant interaction between emotion and age ($F(2,800) = 11.25$, $p < 0.001$), with late MMR values to happy and sad prosodies going more negative with age and late MMR values to angry going more positive with age. Furthermore, the negative-going trend with age was stronger in response to happy than sad prosody. The model is summarized in Table 5, and Figure 8 shows the main effect of electrode region, main effect of sex, and the interaction effect of emotion and age on infants' late MMR amplitudes.

Table 5. Summary of the linear mixed-effect model using the amplitudes of late MMR as the dependent variable.

Factor	*Numerator df*	*Denominator df*	*F*	*p*
Trial-level fixed factors				
Emotion	2	800	12.12	< .001 ***
Region	2	800	19.00	< .001 ***
Laterality	2	799	2.73	0.07
Participant-level fixed factor				
Age	1	32	0.61	0.44
Sex	1	32	5.45	0.03*
Cross-level interaction				
Emotion * Age	2	800	11.25	< .001 ***
Emotion * Sex	2	800	0.87	0.42

* $p < .05$, ** $p < .01$, *** $p < .001$.

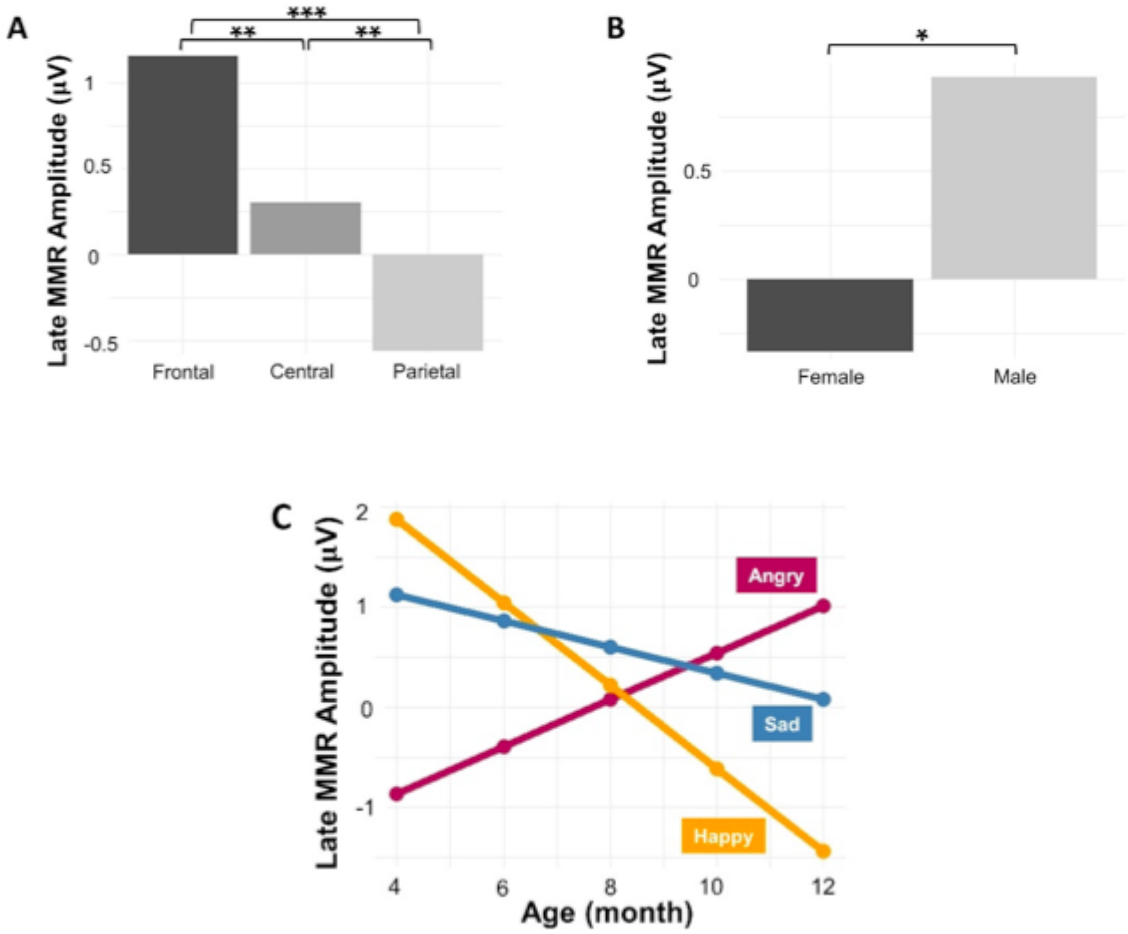


Figure 8. The (A) main effect of electrode region, (B) main effect of infant sex, and (C) interaction effect of emotion and age displayed in model predicted late MMR amplitudes to angry, happy, and sad emotional prosodies.

IV. Discussion

The current study employed a roving multi-feature oddball paradigm to examine infants' neural sensitivities to angry, happy, and sad prosodies against neutral prosody over non-repeating English spoken words. Different vocal emotions elicited distinct MMRs in reference to the neutral voice (Figures 5), indicating that infants can automatically extract the emotional prosodic information across varying words at an early signal processing stage outside attentional focus.

A. Early MMR (100 – 200 ms)

The early MMR is an indicator of listeners' automatic sensory processing of the incoming sounds. It usually shows a positive deflection in early infancy and gradually develops into a more negative-going component (Csibra et al., 2008; Kushnerenko et al., 2013). In the current study, the sad prosody elicited more positive MMR in the early window than angry and happy voices for the infants. We observed an interaction between emotion and age in early MMR that shows a shift from more positive to more negative deflection to happy and sad voices as the age increases, but negative to positive deflection to angry voices. Furthermore, the positive-to-negative trend was more substantial in the sad prosody compared to the happy prosody. For younger infants, sad emotion elicited the most positive early MMR followed by happy emotion, and angry emotion elicited the most negative early MMR. Older infants showed similar MMR to both happy and sad prosodies, but a positive early MMR to angry voices. Infants' distinct early neural responses to happy and angry voices indicate that their pre-attentive system can discriminate between the two high-arousal emotions (Grossmann et al., 2005) that were not distinguishable in their behavioral reactions (Flom & Bahrick, 2007; Walker-Andrews & Lennon, 1991). This differential age effect on angry, happy, and sad voices can be interpreted as the change of infants' neural sensitivities over time. Our data demonstrate that infants before the age of one already distinguish various emotions in human voice at an early neural processing stage, showing a great improvement from the undistinguished early MMR observed in newborns (Kostilainen et al., 2018).

We did not observe a sex effect or interaction between sex and emotion in infants' early MMR. This result is consistent with male and female newborns' similar early MMRs to vocal expressions (Kostilainen et al., 2018). Since early MMR reflects lower-level sensory processing, the lack of the sex effect on this early neurophysiological component re-emphasizes that male and female infants show similar automatic auditory processing of vocal emotional signals. Even though the lack of sex effect is expected in automatic auditory processing, previous infant EEG studies seldom examined the role of biological sex in early sensory processing (McClure, 2000). The current study provides neurophysiological evidence on the similar early processing of emotional speech for male and female infants.

To our knowledge, this is the first report using a multi-feature oddball design to record emotion-modulated MMRs in early infancy. Previous research using a similar experimental design did not observe an emotional effect on newborns' early MMR, even if the vocal emotions carried by simple syllables "ta-ta" should be differentiated more easily (Kostilainen et al., 2018). Our data demonstrated that infants before the age of one already showed neural sensitivities to the acoustic changes of vocal emotions. This automatic processing was more mature for angry and happy prosodies (more negative MMR) than the sad prosody (more positive MMR). Infants' distinct early MMRs to different vocal emotions demonstrate that their auditory system develops rapidly and can process complex vocal emotions efficiently with only a few months of listening experience.

B. Late MMR (300 – 500 ms)

The late MMR is often seen as the infant-version of MMN. Unlike adult MMN, infants' MMR is delayed and emerges as a slow positive wave (Cheour et al., 2000; Leppänen et al., 2004; Trainor, 2010). Over the time of development, infants' slow positive MMR gradually develops into an adult-like negative deflection (see He et al., 2009 for a summary). Compared with the early MMR (100 – 200 ms), previous infant studies on pre-attentive neural responses to emotional prosodies mainly focused on this late MMR (300 – 500 ms) (Cheng et al., 2012; Kostilainen et al., 2020; Kostilainen et al., 2018; Zhang et al., 2014). Some researchers also suggested that this slow positive wave may be a mix of MMR and a following component P3a, a fronto-central oriented positive wave elicited by contextually novel events in adults (Escera et al., 2000; Friedman et al., 2001). Our data confirmed this fronto-centrally oriented late MMR with a positive deflection (Figure 6, B), indicating that the multi-feature oddball task on emotional prosody successfully elicited the target MMR. Our infant listeners showed a stronger, more negative-oriented late MMR to angry than sad voices (especially the younger infants, Figure 5), which is similar to the newborns' data from Cheng et al. (2012) and Zhang et al. (2014). One explanation is that infants' auditory systems are wired to respond to negative, threat-related signals more efficiently, reflected in their stronger late MMR to the angry voice. The novel finding here was that we employed varying spoken words, rather than simple syllables, to deliver emotional prosodies (as in Cheng et al., 2012; Zhang et al., 2014), making the prosodic extraction based on statistical regularities in the acoustic parameters more demanding for the infants. Despite

this challenge, infants in the current study successfully extracted the relevant emotional prosodic categories across changing words and registered the voices differently in their pre-attentive neural system.

An interaction between emotion and age was also observed in this late MMR window. The late happy and sad MMRs shifted from positive to negative deflections with age, but the late angry MMR shifted in the opposite direction. Furthermore, the polarity change was greater to happy and angry prosodies but subtler to the sad prosody. For younger infants, happy emotion elicited the most positive late MMR, and angry emotion elicited the most negative late MMR. In contrast, the older infants showed a more positive late MMR to angry voices and a more negative late MMR to happy voices. Unlike the early MMR, infants showed distinct developmental trends to the three vocal emotions in this later window. The younger group's late MMR pattern was similar to Kostilainen and colleagues' newborn data (2020) that showed more positive happy MMR than angry and sad MMRs. There has been no previous report on older infants' MMRs on the vocal emotions in a multi-feature oddball task. One study measured seven-month-old infants' ERPs to randomly presented angry and happy voices (Grossmann et al., 2005), and the authors observed a more negative ERP to angry than the happy prosody, just as the predicted late MMR amplitudes of 7-month-old infants in our data (Figure 8, C). Even though the two studies used different experimental protocols, it is worth mentioning that Grossmann and colleagues (2005) also used up to 74 different words to deliver each vocal emotion. Together, we are confident that it is practical to use a more complex and natural linguistic context to investigate infants'

neural sensitivities to emotional prosody. Future studies testing older (7 – 12 month olds) infants' MMR to emotional voices are strongly encouraged to scrutinize our current findings.

Unlike the early MMR, we observed a sex effect on infants' late MMR to emotional voices, with male infants showing a more positive late MMR than female infants. Previous studies on newborns' MMR to vocal emotions either did not observe sex differences (Cheng et al., 2012; Kostilainen et al., 2018) or did not include sex as a factor to explain the variations of infants' MMR amplitudes (Grossmann et al., 2005; Kostilainen et al., 2020; Zhang et al., 2014). The similar vocal emotional MMR in male and female newborns (Cheng et al., 2012; Kostilainen et al., 2018) but sex-modulated MMR in three- to 11-month old infants (the current study) suggest that sex differences in emotional voice processing may emerge in the first year of life.

V. Limitations and Future Directions

The current study measures infants' pre-attentive neural sensitivities to emotional prosodies over non-repeating spoken words. While our data suggest that infants successfully extract differences among the four basic emotional prosodies over the varying linguistic context, the results cannot determine if these prosodic categories entail infants' subjective emotional experiences in the way that adults perceive and interpret the emotional prosody categories. This is a common limitation for infant MMR studies, for this particular paradigm only requires participants' automatic auditory detection of different voices without active engagement. To further investigate if emotional evaluation is involved, we may include audiovisual emotional stimuli to test infants'

cross-modal emotional congruency detection and record their EEG signals (Flom & Whiteley, 2014; Grossman, 2013; Otte et al., 2015). Even though including both behavioral and EEG tasks greatly complicates the experimental administration, it ensures that we measure the neurophysiological responses underlying infants' evaluation of emotion.

Another limitation is that our speech stimuli were from one female speaker, limiting the results from being generalized to real-life scenarios where infants listen to the emotional speech from multiple speakers. The rationale of the current study to include non-repeating spoken words from the same female speaker was to establish a more diverse but still manageable linguistic context for infants to extract emotional prosody. Since previous emotional multi-feature oddball tasks for infants mainly used a few simple and fixed syllables (Cheng et al., 2012; Kostilainen et al., 2020; Kostilainen et al., 2018; Zhang et al., 2014), it is reasonable to first examine their emotional MMRs to varying words from the same speaker before moving on to multiple speakers.

Despite the limitations, our results and Grossmann and colleagues' (2005) report confirm that infants' pre-attentive neural systems can group words with the same emotional prosody against other words delivering a different emotional prosody. Future studies can start adding speech stimuli from male speakers to create an even more natural listening context and thoroughly examine the sex effect on infants' early processing of emotional voices.

More importantly, the results from our efficient passive-listening multi-feature oddball paradigm have great implications for future studies investigating emotional

voice processing in populations with short sustained attention to lengthy tasks. For instance, Korpilahti et al. (2007) studied the differences in neural responses to angry voices between children with and without Asperger syndrome. By incorporating the multi-feature design, researchers can examine more emotional voices and understand their different effects on speech processing in neurodivergent infants and children.

VI. Conclusion

The current study establishes the feasibility of the multi-feature oddball paradigm in studying early emotional prosody speech perception by successfully eliciting infants' early and late MMRs to happy, angry, and sad prosodies using non-repeating spoken words. The results clearly revealed distinct developmental changes in infants' neural activity patterns to each emotional category, indicating that EEG is a sensitive tool that captures developmental trends that may be obscure in behavioral studies. Finally, we observed different MMR amplitudes in male and female infants in the late but not early MMR window. Since this sex effect is not observed in neonates' MMR, we may infer that sex differences in the neural correlates of emotional speech emerge after infants gain some listening experience. Further research is required to determine the role of biological and social factors in the commonly observed sex differences in processing socio-emotional signals and the functional significance of age and emotion category interaction effects in the early and late MMRs in language learning and socio-emotional development.

Chapter 4: Gender Differences Revealed Outside the Focus of Attention to Emotional Prosody Variation in Spoken Words (Study 3)

This chapter has been submitted to the Journal of Speech, Language and Hearing Research.

I. Introduction

Daily communication seldom consists of neutral speech. Speakers express their views through both content (i.e., what is said) and style (i.e., how it is said) of the speech, and listeners need to evaluate both cues to fully understand the message properly. Emotional prosody is one of the speaking styles that speakers use to display their internal states through varying pitch, intensity, stress, and temporal information in the voice (Banse & Scherer, 1996). The same sentence can carry a very different message once the speaker changes the emotional intonation. In the case that semantic meaning contradicts the prosody in the voice, listeners tend to rely more on the prosodic information (Ben-David et al., 2016; Filippi et al., 2017; Kim & Sumner, 2017; Lin et al., 2020; Mehrabian & Wiener, 1967; Schirmer & Kotz, 2003). Therefore, timely processing of emotional speech prosody is essential in daily communication.

To capture listeners' online tracking and perception of the fast-changing speech prosodic information, the time-sensitive measurement—electroencephalography (EEG)—can be used. Previous EEG studies on emotional processing focused more on the visual than the auditory modality (Grossmann et al., 2005; Thierry & Roberts, 2007). Among the few reports on listeners' event-related potentials (ERPs) to emotional prosody, which typically require hundreds of trials for averaging the time-locked EEG responses, the speech stimuli have used simple contrasts of vowels (Carminati et al.,

2018), simple syllables (Fan et al., 2013; Hung & Cheng, 2014; Schirmer, Striano, et al., 2005), or words (Jiang et al., 2014; Thönnessen et al., 2010; Zora et al., 2020) to carry the vocal emotions in fixed and repeated trials. However, it remains unclear how listeners' brains register the emotional prosodic category across highly varying linguistic carriers, which better resembles the diverse listening environment in real life. Therefore, this current study aimed to examine listeners' early neural responses to natural emotional prosodies in non-repeating spoken words in a roving stimulus presentation paradigm that does not require attentive listening, supplementing the literature on vocal emotion perception with more restricted speech stimuli. Furthermore, instead of a simple contrast, we included three vocal expressions of emotion—angry, happy, and sad—against neutral prosody in a single EEG recording session to investigate whether listeners would show distinguishable neural activities in extracting each emotional prosody from the roving stimulus presentation. Successful establishment of this protocol can benefit future emotional prosody research with populations such as infants and children, who have relatively short attentional span, to inspect their neural responses to multiple emotions in voices within one EEG recording session.

Multi-Feature Oddball Paradigm for Emotional Prosody Speech Perception

The auditory oddball paradigm is the most used task for recording early pre-attentive neural responses to assess neural sensitivity in how the central auditory system automatically discriminates differences in speech and nonspeech stimuli (for a review, see Näätänen et al., 2007). In a typical implementation of this paradigm, one sound is repetitively presented 80 – 85% of the time (i.e., the *Standard*), and this stream of sound

is randomly interrupted by another sound that is presented 15 – 20% of the time (i.e., the *Deviant*). The event-related potentials (ERPs) to the *Standard* and *Deviant* are compared, yielding a difference ERP (the *Standard* ERP subtracted from the *Deviant* ERP) that denotes listeners' neural discriminatory response to the two sounds (*Standard* and *Deviant*). Since the paradigm does not require listeners' voluntary attention, the difference ERPs are usually interpreted as listeners' pre-attentive detection or automatic processing of the *Deviant* sound in direct comparison with the sensory-memory trace built on the stream of *Standard* sound. For instance, researchers can assign one emotional voice as the *Standard* and another as the *Deviant* and analyze the difference ERPs to index listeners' pre-attentive neural sensitivities to the two emotional prosodies.

Two ERP components—the mismatch negativity (MMN) and P3a—are commonly observed in the difference ERPs to an emotional prosodic change in speech (Carminati et al., 2018; Hung & Cheng, 2014; Pakarinen et al., 2014; Wambacq & Jerger, 2004; Zora et al., 2020). The MMN response typically peaks at approximately 150~200 ms after the onset of acoustic change in the *Deviant* relative to the *Standard* stimuli, and it appears as a negative deflection in the difference ERPs at centro-frontal electrodes over the scalp (e.g., Fan et al., 2013; Jiang et al., 2014; Schirmer, Striano, et al., 2005; Thönnessen et al., 2010). The MMN amplitude tends to be larger with perceptually more distinct *Standard* and *Deviant* stimuli, and it is interpreted as a sensitivity index of listeners' perception of the two auditory inputs (Garrido et al., 2009; Näätänen et al., 2007). Even though the MMN was initially linked to low-level acoustic processing, many studies showed stronger MMN amplitudes to prosodic change in real

words than pseudowords (Fan et al., 2013; Zora et al., 2020), indicating some degrees of higher-level cognitive processing at this early stage. Indeed, studies testing natural emotional prosody using syllables (Hung & Cheng, 2014; Schirmer & Escoffier, 2010; Schirmer, Striano, et al., 2005) and complex words (a set of 16 words, Jiang et al., 2014) successfully recorded MMN to emotional prosodic change from neutral to happy, fearful, or angry, demonstrating that listeners' pre-attentive system can capture the differences in higher-level emotional prosodic categories. These reports indicate that the MMN can be a reliable neurophysiological measure to examine listeners' neural sensitivity to affective prosodic categories.

Following MMN, P3a is a positive deflection elicited around 350 ms after the emotional sound onset, and it is usually fronto-centrally oriented over the scalp (Goydke et al., 2004; Hung & Cheng, 2014; Wambacq & Jerger, 2004; Zora et al., 2020). Unlike MMN that is associated with both acoustic- and cognitive-level processing, P3a is mainly linked to cognitive evaluations of the incoming sounds and the involuntary attention switch to the novel auditory input (Escera et al., 2000; Escera et al., 1998; Escera et al., 2001; Näätänen et al., 2007; Polich, 2007). The P3a component is especially sensitive to emotional prosodic information, such that voice changes from neutral to affective prosody consistently elicited stronger P3a response (Carminati et al., 2018; Pakarinen et al., 2014). Moreover, Zora et al. (2020) measured listeners' P3a to both emotional and non-emotional prosody (i.e., word stress) to examine if P3a is sensitive to any prosodic information. They found that P3a amplitudes were stronger to the emotional prosody than non-emotional prosody, indicating that the elicitation of P3a

is related to the affective salience of the auditory context, not just the acoustic salience of the speech prosody.

To date, the findings on the MMN and P3a components have mainly tested two emotional prosodies (one *Standard* and one *Deviant*) in a single session. A systematic assessment of neural sensitivities to multiple vocal emotional categories has rarely been examined (except Carminati et al., 2018). Furthermore, researchers pointed to the concerns that emotional prosodies were delivered through a small number of fixed syllables or words, limiting the generalization of natural emotional voice processing at the neural level (Zora et al., 2020). To compare the pre-attentive neural responses to more emotional categories, we turned to the multi-feature oddball paradigm (or optimal paradigm, Näätänen et al., 2004; Pakarinen et al., 2009). As a modified auditory oddball task, the multi-feature oddball paradigm limits the presentation of the *Standard* sound to 50% and allows different types of *Deviants* to equally take up the rest of the 50% sound presentation. As a trade-off, the differences among ERPs elicited by multiple *Deviants* can be subtle and require a more sophisticated statistical modeling approach than the traditional oddball task (at least 80 % of *Standard* sound). To our knowledge, previous research has not employed the multi-feature oddball paradigm to examine pre-attentive neural responses to emotional prosody over non-repeating spoken words (only one study with 14 pseudo-words, Thönnessen et al., 2010). Therefore, the first aim of the present study was to test the feasibility of the protocol, i.e., whether including three vocal emotions as three *Deviants* in a multi-feature oddball task can successfully elicit the MMN and P3a responses, the two neural markers for auditory emotional change

detection. To address the issue of generalizability of natural emotional prosody, we included non-repeating spoken words to deliver each emotional prosody (see details in the Method section). This way, the presence of MMN and P3a would reflect the categorization of the suprasegmental prosodic information from the constantly changing spoken words, not the specific acoustic change from emotions embedded in the same lexical items.

Gender Effect on Emotional Prosody Perception and the MMN and P3a Components

Gender differences have been observed in emotional prosody speech perception (Hall, 1978; Schirmer et al., 2002; Schirmer, Kotz, et al., 2005; Schmid et al., 2011; Sen et al., 2018; Thompson & Voyer, 2014), with more reports on women's higher sensitivities to emotional information in human voice (e.g., Demenescu et al., 2014; Paulmann et al., 2008). For instance, female listeners recognize subtle emotional tones better, and they are more susceptible to conflict information from the prosodic domain (Lin, in press).

While reports based on behavioral responses reflect a relatively late processing stage involving decision-making for each test trial, neurophysiological measurements can examine listeners' automatic processing and involuntary attention/orienting to the auditory signals even before making any behavioral decision. Furthermore, neurophysiological measurements may reveal the refined time scale of differential attention allocation while listeners show similar behavioral responses, making the auditory EEG study a great tool to investigate potential gender effects at early emotional

speech processing outside attentional focus. In previous studies using the auditory oddball paradigm, female listeners showed stronger MMN to emotional prosody change than male listeners (Fan et al., 2013; Schirmer, Striano, et al., 2005), which may indicate women's higher neural sensitivity to the acoustic features of emotional voices. To further determine if women's stronger involuntary neural response is solely elicited by the acoustic change, some reports included acoustic controls to remove the speech context and found no gender differences in the MMN response (Fan et al., 2013; Hung & Cheng, 2014; Nagy et al., 2003). Taken together, emotional information may facilitate female listeners' auditory change detection at an early auditory processing stage before they start allocating attentional resources to the auditory inputs.

Gender effects on P3a response to vocal expressions of emotion have seldom been reported. Hung and Cheng (2014) observed a larger P3a to emotional prosody change in women than men, but there were no emotion-dependent gender differences. Another study used visual distractors to induce emotional context while listeners were passively listening to emotional voices, and found that only female listeners' P3a amplitudes were affected by the emotional context (Garcia-Garcia et al., 2008). Given that the P3a indexes more attention-orientation to the incoming sounds, it is reasonable to observe gender effects on the P3a component that may underlie the different performances in emotional prosody recognition tasks across gender groups. It could be the case that the simple dichotomic setup of emotional prosody contrast using fixed linguistic items in previous studies obliviated potential gender differences in involuntary attentional orienting towards prosodic changes.

The Current Study

The goal of the current study was two-fold. First, we examined whether the MMN and P3a components would be elicited by three emotional prosodic *Deviants* (angry, happy, and sad) in a multi-feature oddball task with non-repeating spoken words. The use of different spoken words as opposed to fixed repeated syllables (e.g., Hung & Cheng, 2014; Schirmer & Escoffier, 2010; Schirmer, Striano, et al., 2005) or limited numbers of pseudowords (e.g., Frühholz et al., 2011) enforces listeners to extract paralinguistic category across varying lexical item contents. Even though the non-repeating lexical contents create a complex acoustic context for listeners, we expected that listeners would still build their auditory memory trace based on the emotional prosodic category and show both MMN and P3a responses. As previous reports observed the strongest MMN and P3a to high-arousal negative emotions (Carminati et al., 2018; Hung & Cheng, 2014), we further expected to see the strongest MMN and P3a to the angry *Deviant*. Second, we investigated the gender effects on listeners' early (i.e., MMN) and late (i.e., P3a) involuntary neural processing of emotional prosodic change. Based on the previous EEG reports, the gender effect may be emotion-specific. For instance, one report found that women showed stronger MMN than men to fearful voices, but both showed similar MMNs to happy voices (Hung & Cheng, 2014). In the same report, women showed stronger P3a regardless of the emotional category. Due to the scarcity of empirical studies on this topic, making highly specific hypotheses for each emotional voice may be overstretched. Building on our first hypothesis and previous research (Schirmer, Striano, et al., 2005), we expected that our exploratory

analysis would demonstrate the "female advantages" with stronger MMN and P3a to some (if not all) of the emotional prosodic changes.

II. Method

A. Participants

The participants were 22 monolingual native speakers of American English studying at the University of Minnesota. All participants (female = 11, male = 11) were right-handed, aged between 18 and 28 (mean = 20.8), and without hearing- and language-related problems. They all had normal or corrected-to-normal vision. The experimental protocol was approved by the Institutional Review Board at the University of Minnesota. Participants signed the informed consents before the experiment, and each received $10 upon completion.

B. Stimuli

All speech stimuli were taken from the Toronto Emotional Speech Set (TESS, Dupuis & Pichora-Fuller, 2010), which includes 200 monosyllabic phonetically balanced words (Northwestern University Auditory Test No. 6, NU-6; Tillman & Carhart, 1966) as listed in Appendix B. Each of the 200 words was spoken in neutral, happy, sad, and angry voices by a young female speaker, yielding a total of 800 stimuli. The sounds were sampled at 24414 Hz, with the mean sound intensity levels equalized using Praat 6.0.40 (Boersma & Weenink, 2020). Table 1 summarizes the mean fundamental frequency (F0), duration, intensity variation, harmonics-to-noise ratio (HNR), and spectral centroid in each emotional prosody. These five acoustic measures

are commonly used to characterize different vocal emotions (Amorim et al., 2019; Banse & Scherer, 1996; Johnstone & Scherer, 2000; Mani & Pätzold, 2016).

Table 6. *The acoustic properties of each emotional prosody*

Emotions	*Mean F0 (Hz)*	*Duration (ms)*	*Intensity Variation (dB)*	*HNR (dB)*	*Spectral centroid (Hz)*
Angry	216.71	646	11.15	9.22	1810.96
Happy	226.13	742	10.82	17.53	1052.92
Sad	180.42	822	10.18	19.31	408.79
Neutral	195.04	667	9.14	18.75	758.43

Note. The averaged values of all the words were used to report the mean fundamental frequency (F0), word duration, intensity variation, harmonics-to-noise ratio (HNR), and spectral centroid in each emotional prosody.

C. Procedure

We adopted a multi-feature oddball paradigm (or optimal paradigm, Näätänen et al., 2004; Pakarinen et al., 2009; Thönnessen et al., 2010) to examine listeners' early neural sensitivities to happy, angry, and sad prosodies as against the neutral prosody. The multi-feature oddball paradigm is a passive listening protocol, and it allows us to measure three emotional prosody contrasts (from neutral to angry, from neutral to happy, and from neutral to sad) within the same recording session. We presented 600 trials in total. The *Standard* stimuli were words in a neutral tone (presented with 50% probability, 300 trials). The three emotional tones (happy, angry, sad) served as three types of *Deviant* stimuli (each presented with 16.7% probability, 100 trials). For the 300 *Standard* trials, all the 200 words in neutral voice were used, and 100 words were randomly selected for repetition. For each type of the 100 *Deviant* trials, 100 words in

each emotional voice were randomly selected and used. The sounds were always presented in alternating *Standard* and *Deviant* fashion. The three types of *Deviant* (three emotions) were pseudo-randomly interspersed, with no consecutive *Deviant* trials in the same emotional prosody (see Figure 9 for an example of the sound presentation order). The inter-stimulus interval (ISI) was randomized between 800 – 900 ms, and the total recording time was around 25 minutes.

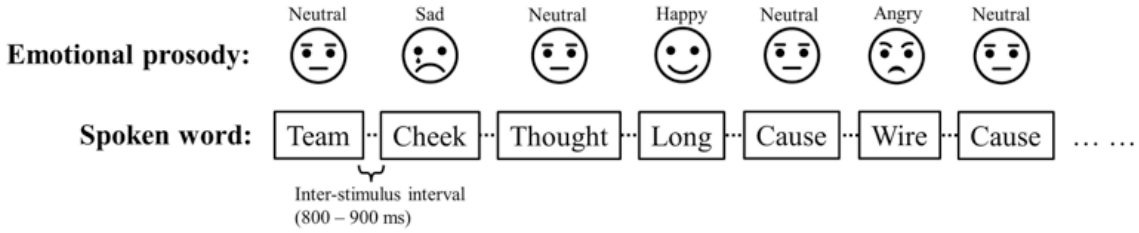


Figure 9. A schematic example of the order of the trials. The *Standard* (neutral prosody) and *Deviant* (angry, happy, and sad prosodies) were always alternating, and the three emotions (*Deviants*) were pseudo-randomly interspersed.

Participants were seated in an electrically and acoustically treated booth (ETS-Lindgren Acoustic Systems) with a 64-channel WaveGuard EEG cap. They were instructed to ignore the speech sounds and focus on a silent movie while the continuous EEG signals were recorded. The speech sounds were played via two loudspeakers (M-audio BX8a) placed at a 45-degree azimuth angle 3 feet away from the participants and presented at 55 dB SL relative to the individual listener's hearing threshold at 1 kHz (Koerner & Zhang, 2015). The sound presentation was controlled by E-Prime (Psychological Software Tools, Inc) using a Dell PC outside the sound-treated room. Continuous EEG data were recorded through the Advanced Neuro Technology EEG System (Advanced Source Analysis version 4.7). The WaveGuard EEG cap has a layout

of 64 Ag/AgCl electrodes following the standard International 10-20 Montage system with intermediate locations, and it is connected to a REFA-72 amplifier (TMS International BV). The default bandpass filter for raw data recording was set between 0.016 Hz to 200 Hz, and the sampling rate was 512 Hz. The electrode AFz served as the ground electrode. The impedance of all electrodes was kept under 5 kΩ.

D. Data analysis

The continuous EEG data preprocessing was completed offline by EEGLAB v14.1.1 (Delorme & Makeig, 2004). The continuous EEG data were low-pass filtered at 30 Hz, downsampled to 250 Hz, and high-pass filtered at 0.5 Hz. The EEG data were then re-referenced to the average of the two mastoid electrodes. Next, we applied the "Clean_rawdata" EEGLAB plug-in to remove low-frequency drifts and non-brain activities (e.g., muscle activity, sensor motion, etc.). Data were then decomposed by the Independent Component Analysis (ICA) algorithm (Dammers et al., 2008; Delorme et al., 2001) to remove eye-blink artifacts. ERP epochs were extracted from 100 ms pre-stimulus onset[6] to 1000 ms post-stimulus onset, and baseline correction was applied using the mean voltages of the 100-ms baseline period. Epochs containing data points over the range of 100.0 μV were rejected before averaging. The numbers of trials that remained for each emotional prosody were 276 for neutral (*Standard*), 89 for happy (*Deviant*), 91 for angry (*Deviant*), and 91 for sad (*Deviant*). Using ERPLAB v7.0.0 (Lopez-Calderon & Luck, 2014), averaged event-related potentials (ERPs) were derived for *Standard* (neutral prosody) and each three types of the *Deviant* (angry, happy, and

[6] A comparison of ERP epochs extracted relative to the sound onset or vowel onset and the statistical results are included in Appendix C.

sad prosodies). Difference waveforms were created by subtracting the *Standard* ERP from each *Deviant* ERP, yielding happy, angry, and sad difference waveforms.

All statistical analyses were completed in R (https://www.r-project.org/) with the packages "lme4" (Bates et al., 2015), "lmerTest" (Kuznetsova et al., 2017), and "emmeans" (Lenth et al., 2018). The difference waveforms were used for assessing the target components MMN (200 – 300 ms) and P3a (350 – 450 ms). The time window for each component was selected based on previous neurophysiological reports of emotional prosody perception (Pakarinen et al., 2014; Thönnessen et al., 2010; Zora et al., 2020) and visual inspection of the current grand average difference waveforms. The amplitudes of MMN and P3a for statistical analyses were calculated as the mean voltages of the 40 ms peak (20 ms before and after the peak value) of the difference waveforms within the two time windows (MMN, 200 – 300 ms; P3a, 350 – 450 ms). The most negative 40-ms peak value was extracted for MMN, and the most positive 40-ms peak for P3a. Peak amplitude extraction was applied to channels at frontal (F-line, F3, Fz, F4), central (C-line, C3, Cz, C4), and parietal (P-line, P3, Pz, P4) regions (Zora et al., 2020). These peak amplitudes were then used as the dependent variables in the later statistical models.

Linear mixed-effect models were respectively implemented on MMN and P3a amplitudes. Each model included by-participant intercept as a random-effect factor. Deviant/Emotion (happy, sad, and angry), region of the electrode (anterior, central, and parietal), and laterality of the electrode (left, middle, and right) were included as trial-level fixed-effect factors. Gender (female and male) was included as a participant-level

fixed factor. Finally, cross-level interactions of emotion and gender were also included. Post-hoc t-tests with Bonferroni corrections ($\alpha = .05$) were performed to characterize any significant interaction.

III. Results

The linear mixed-effect models allowed an in-depth analysis of potential contributors to the MMN and P3a amplitude measures extracted from the difference waveforms, derived from subtracting the *Standard* ERP from the *Deviant* ERPs. Both female and male listeners showed distinct MMN and P3a peaks to the change of emotional prosody. Their grand mean ERP waveforms and difference waveforms recorded from midline electrodes (Fz, Cz, Pz) for all emotional prosodies are displayed in Figure 10. The topographic maps of MMN and P3a peaks of each emotional prosody are presented in Figure 11.

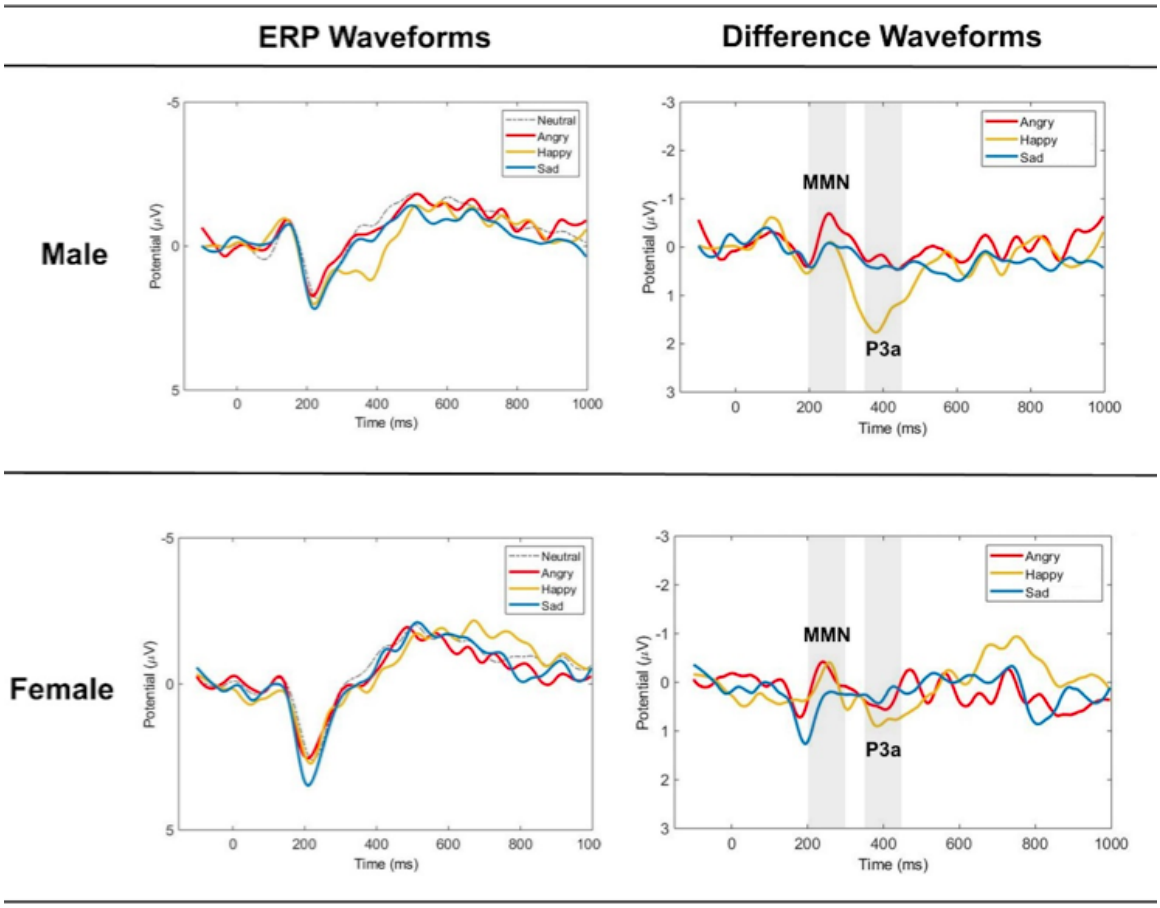


Figure 10. The grand mean event-related potential (ERP) waveforms of *Standard* (neutral prosody) and *Deviants* (angry, happy, and sad), and grand mean difference waveforms of angry, happy, and sad for male and female listeners. Mean amplitudes of the midline electrodes (Fz, Cz, Pz) were used for the waveforms. The gray shaded areas mark the windows for MMN (200 – 300 ms) and P3a (350 – 450 ms).

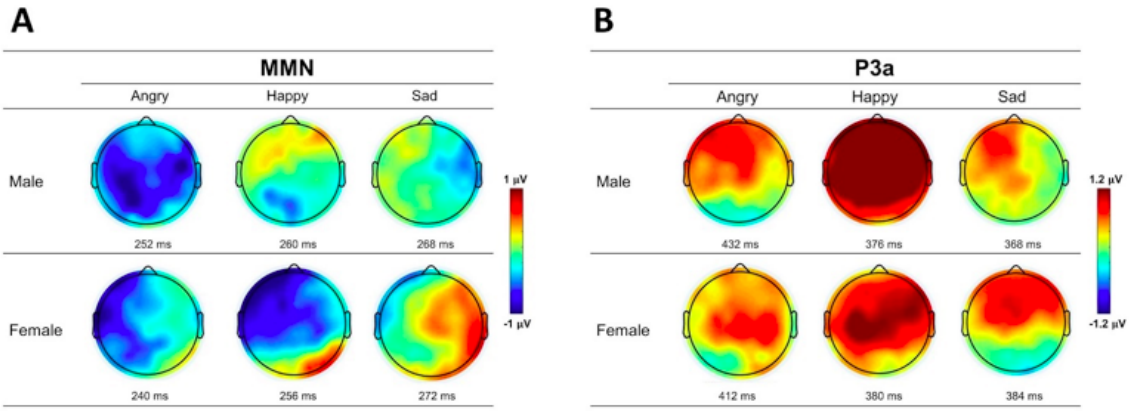


Figure 11. The scalp topographic maps of (A) MMN and (B) P3a to angry, happy, and sad emotional prosodies averaged across male and female listeners. The topographies are based on the latencies of peak values at Cz channel.

A. Mismatch Negativity (MMN)

The main effect of emotion ($F(2,572) = 8.21$, $p < 0.001$) was significant, but not the main effects of electrode region ($F(2,572) = 1.27$, $p = 0.28$) or electrode laterality ($F(2,572) = 0.59$, $p = 0.55$). In general, MMN to angry prosody was stronger than happy ($p = 0.002$) and sad prosodies ($p = 0.001$). The participant-level main effect of gender was not significant ($F(1,22) = 0.0013$, $p = 0.97$), but there was a significant interaction between emotion and gender ($F(2,572) = 4.68$, $p = 0.009$). Post-hoc t-tests revealed that male listeners showed stronger MMN to angry than happy prosody ($p = 0.002$), whereas female listeners showed stronger MMN to angry than sad prosody ($p = 0.02$). The model results are summarized in Table 7, and Figure 12 shows the interaction effect of emotion and gender on MMN amplitudes.

Table 7. Summary of the linear mixed-effect model using the amplitudes of MMN as the dependent variable.

Factor	*Numerator df*	*Denominator df*	*F*	*p*
Trial-level fixed factors				
Emotion	2	572	8.21	< .001 ***
Region	2	572	1.27	0.28
Laterality	2	572	0.59	0.55
Participant-level fixed factor				
Gender	1	22	0.001	0.97
Cross-level interaction				
Emotion * Gender	2	572	4.68	.009 **

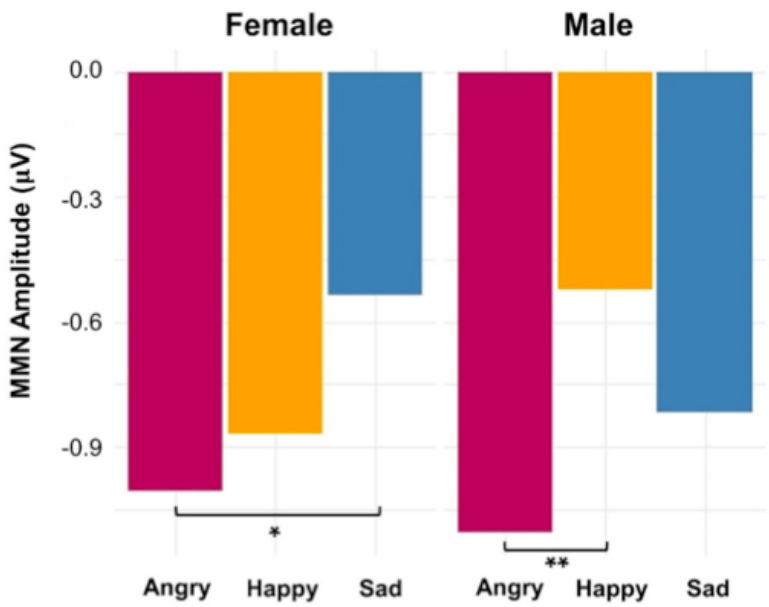


Figure 12. The interaction effect of emotion and gender displayed in the model predicted MMN amplitudes to angry, happy, and sad emotional prosodies in male and female listeners.

B. P3a Component

The main effects of emotion ($F(2,572) = 25.67$, $p < 0.001$) and electrode region ($F(2,572) = 26.13$, $p < 0.001$) were significant, but not the main effect of electrode laterality ($F(2,572) = 1.54$, $p = 0.21$). In general, P3a to the happy prosody was stronger than angry and sad prosodies (*ps* < 0.001), and frontal channels and central channels recorded stronger P3a than parietal channels (*ps* < 0.001). The participant-level main effect of gender was not significant ($F(1,22) = 0.49$, $p = 0.49$), but there was a significant interaction between emotion and gender ($F(2,572) = 14.61$, $p < 0.001$). Post-hoc t-tests revealed that male listeners showed stronger P3a to happy than angry ($p < 0.001$) and sad ($p < 0.001$) prosodies, whereas female listeners showed similar P3a to all

emotional prosodies. The model results are summarized in Table 8, and Figure 13 shows the main effect of electrode region and the interaction effect of emotion and gender on P3a amplitudes.

Table 8. Summary of the linear mixed-effect model using the amplitudes of P3a as the dependent variable.

Factor	*Numerator df*	*Denominator df*	*F*	*p*
Trial-level fixed factors				
Emotion	2	572	25.67	< .001 ***
Region	2	572	26.13	< .001 ***
Laterality	2	572	1.54	0.21
Participant-level fixed factor				
Gender	1	22	0.49	0.49
Cross-level interaction				
Emotion * Gender	2	572	14.61	< .001 ***

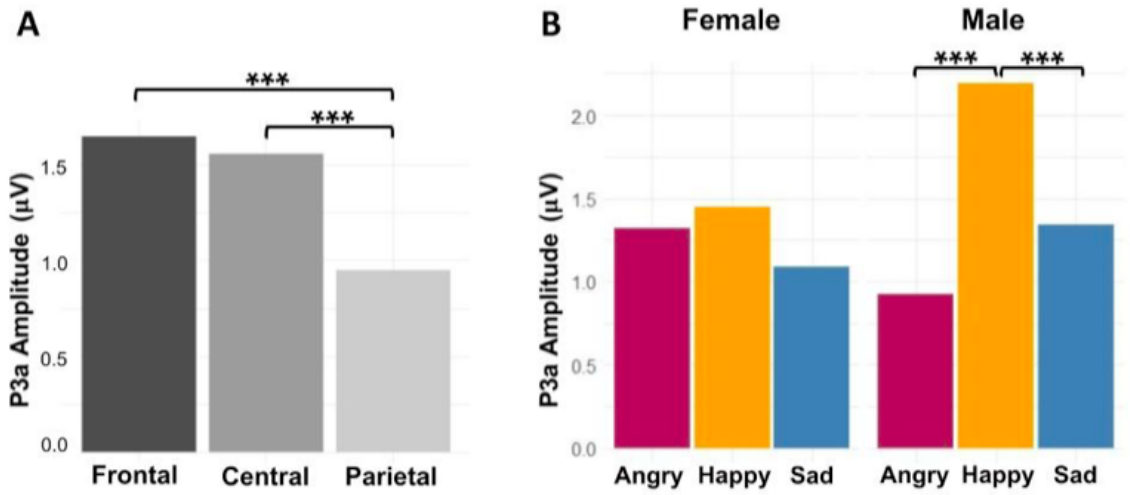


Figure 13. The (A) main effect of electrode region and (B) interaction effect of emotion and gender displayed in the model predicted P3a amplitudes to angry, happy, and sad emotional prosodies in male and female listeners.

IV. Discussion

Emotional prosody in natural speech is a crucial social cue that listeners need to detect efficiently for effective interpersonal communication, but it has not received as much attention as the visual emotional expression has. In this study, we employed a multi-feature auditory oddball paradigm to examine listeners' early and late involuntary neural responses to emotional prosody change. Unlike previous studies using limited lexical contents to present emotional voices, we adopted roving stimulus presentation with varying spoken words to deliver natural emotional prosodies that are usually embedded in complex language contexts in real life. Listeners in our study successfully extracted emotional prosodic information from non-repeating spoken words by showing MMN and P3a to the three emotional *Deviants*—angry, happy, and sad voices. The elicitation and confirmation of these two components at the individual as well as group level from our modified multi-feature oddball task demonstrated that abstract categories of emotional prosodic information could be tested in naturally changing speech stimuli even with high acoustic variations in the linguistic domain. More importantly, the time-efficient design of testing three emotional *Deviants* in one task did not appear to compromise the target ERP components for emotional prosodic change detection. Our results also corroborate the findings that the MMN response reflects not only the simple detection of acoustic change but also the change at higher-level category (Jiang et al., 2014; Näätänen et al., 2001; Picton et al., 2000; Zora et al., 2020). Consistent with our

predictions, emotion-specific gender effects were found in both the early (i.e., MMN) and late (i.e., P3a) neural sensitivity markers.

A. MMN—Angry Voices Elicited the Strongest Response in Both Men and Women

A stronger MMN response was observed when the background voice changed from a neutral tone to an angry tone than changes to other emotional categories. This stronger MMN activity reflected heightened automatic processing of the ambient high-arousal negative emotional sounds, even if the listeners were not paying attention to the auditory events. The early time window of MMN (around 200 ms after sound onset) also implies that the pre-attentive sensory processing is activated early on when the affective signals change. The early enhanced response to emotions such as anger or fear, so-called "negative bias," is considered essential for survival because these sounds are usually associated with immediate threat or danger (Adolphs, 2002; Scherer, 1989; Schirmer, Striano, et al., 2005). Our results align with the notion of negativity bias response, and similar results have been shown in an earlier EEG study (Carminati et al., 2018). Collectively, the data suggest that the human pre-attentive system employs a fast and automatic check of the incoming sounds and responds quickly to category contrasts of affective voices, particularly the voices that signal potential threats.

Clear gender differences were observed in the ERP measures. Female listeners showed stronger MMNs to angry voices than sad voices, and male listeners showed stronger MMNs to angry voices than happy voices. In other words, women's early neural response distinguishes angry from sad voice but not from happy voice, whereas men's early neural response distinguishes angry from happy voice but not from sad

voice. Female listeners' MMN was not stronger than male listeners' in any emotional *Deviants*, which differs from the previous findings that generally show enhanced emotional MMN in women than men (Fan et al., 2013; Hung & Cheng, 2014). From the waveforms plot (Figure 10), women's indistinguishable MMN to angry and happy sounds was mainly related to their enhanced MMN for the happy prosody. Female listeners' higher sensitivity to happy prosody corresponds with their greater automatic processing of happy facial expressions (Donges et al., 2012), indicating their perceptual advantage of positive emotion processing even at the pre-attentive level. On the other hand, male listeners showed a more distinct MMN to angry prosody than the other two emotions in Figure 10. Our results contrast with the view that men usually show weaker emotional MMNs; instead, the data indicate differential response patterns in male listeners for various basic categories of emotional prosody with heightened early neural sensitivity to the angry voice. Here, the preserved pre-attentive processing sensitivity to angry prosody in our male participants is consistent with previous behavioral studies showing their heightened response to angry emotional information (Kret & De Gelder, 2012).

While a stronger MMN response to angry voice detection was elicited in all listeners, there appear to be subtle gender effects even at the pre-attentive level. Thus, gender-specific adaptive strategies might play an important role in how males and females automatically extract emotional prosody information and respond to the changes in the surrounding affective auditory signals.

B. P3a—Happy Voices Elicited the Strongest Response, Especially in Men

The P3a response measured in the current study was fronto-centrally oriented, consistent with the topographic distribution of the classical P3a (Polich, 2007; SanMiguel et al., 2010). Statistical results confirmed the significant main effect of the electrode region. As a later involuntary neural response following the MMN, the P3a reflects listeners' involuntary attentional shift to the novel auditory input in the background and involves some signal appraisal. Among the three emotional *Deviants*, our listeners (particularly the male listeners) showed the strongest P3a to the happy voice. Previous reports mainly focused on the enhanced P3a component to general affective information but seldom inspected P3a differences for each emotional prosody (Jiang et al., 2014; Pakarinen et al., 2014; Thönnessen et al., 2010; Zora et al., 2020). One study by Pinheiro et al. (2017) used laughter and growl to present happy and angry voices, and they asked participants to pay attention to the sounds during the EEG recording. Their results showed enhanced positive deflection to laughter at 350 – 450 ms after the sound onset, similar to the time window of our P3a component. Another report presented different emotional prosodies over French vowels and observed stronger P3a to the happy voice than sad and neutral voices (Carminati et al., 2018). Along with our results showing increased P3a to happy prosody than angry and sad prosodies over non-repeating spoken words, listeners may involuntarily orient their attention to positive information more than negative ones after the initial sensory processing stage (i.e., MMN).

Gender differences were also present in the P3a response. The P3a effect for the happy voices was primarily driven by the male, not female listeners. In other words, women did not show distinguishable P3a responses to the three emotional *Deviants*. P3a is a component that can be easily habituated, which means that its amplitude declines as the listeners are more experienced with the *Deviant* events (Friedman et al., 2001). It is possible that arousal and involuntary orienting to affective signals in female participants show similar habituation or saturation effects across the different emotional prosody categories, rendering indistinguishable P3a responses to the emotional prosodic changes. As most of the studies on P3a to emotional prosody change did not examine gender effects (Carminati et al., 2018; Pakarinen et al., 2014; Zora et al., 2020), more research is needed to establish the link between P3a and its functional significance associated with listeners' subsequent behavioral actions to better interpret the gender differences in this late neural sensitivity marker to emotional prosody perception.

V. Limitations and Future Directions

The current study aimed to employ a roving multi-feature oddball paradigm to examine listeners' pre-attentive neural sensitivities to emotional prosody in speech. We only included speech stimuli from female speakers because the emotional speech set of phonetically balanced words only contains female-voice recordings (Dupuis & Pichora-Fuller, 2010). One neurophysiological study demonstrated that listeners showed early neuro-differentiation of emotional prosody information regardless of the speakers' gender (Paulmann & Kotz, 2008). However, a recent behavioral study observed a modulatory effect of encoder gender of the speech stimuli on listeners' emotional

prosody recognition (Lin et al., in press). In this regard, including both female and male emotional voices in the stimuli can provide a more fine-grained view on listeners' neural sensitivities to natural emotional speech prosody and the potential gender differences that may be influenced by the gender of the speaker.

Second, we incorporated non-repeating real words to create a more natural linguistic context for delivering emotional prosody. Even though we carefully selected a phonetically-balanced word list to control phonetic-level acoustic variations across emotional voices, the paralinguistic features such as pitch, intensity variation, or word durations still co-vary with different emotional prosodies. Singling out each acoustic feature in emotional voices and testing each of them may not be realistic, because emotional prosody is essentially a collective of all the relevant acoustic properties (Bachorowski & Owren, 2008; Banse & Scherer, 1996; Johnstone & Scherer, 2000). One solution is to create four oddball tasks and use each of the neutral, happy, angry, and sad prosodies as the *Standard* sound, and compare *Standard* and *Deviant* sounds of the same emotion across tasks. This solution may not be the most optimal one because it contradicts our purpose to establish an efficient testing protocol to record MMN and P3a to multiple emotional *Deviants* that can potentially be applied to clinical and pediatric populations without requiring focused attention and extended hours of EEG recording. Nonetheless, a follow-up study with several multi-oddball recording sessions will still be valuable to verify the findings about the MMN and P3a components to the three emotional voices measured in the current study. Our roving multi-feature oddball protocol and findings add to the existing literature on neural sensitivities to emotional

prosody using a wide range of lexical items such as vowels (Carminati et al., 2018), simple syllables (Fan et al., 2013; Hung & Cheng, 2014; Pakarinen et al., 2014; Schirmer & Escoffier, 2010; Schirmer, Striano, et al., 2005), limited numbers of words (Jiang et al., 2014; Thönnessen et al., 2010), or non-speech sounds (Thierry & Roberts, 2007). Collectively, these data not only help establish the feasibility of the neurophysiological approach but also provide in-depth evidence on how human pre-attentive system captures the change of the incoming emotional prosody change in speech and how the involuntary attentional system is triggered in early stages of emotional prosody processing, including gender differences, which has important implications for future developmental and clinical studies (Charpentier et al., 2018; Paris et al., 2018; Zhang et al., 2021).

VI. Conclusion

Using a passive listening paradigm with EEG recording, we assessed adult listeners' pre-attentive neural sensitivity in extracting and discriminating affective prosodic categories across roving stimuli of spoken words and the following involuntary orientation to prosodic contrasts without overt behavioral reactions. The MMN and P3a results not only demonstrated the feasibility of our roving multi-feature oddball task but also revealed important gender differences in emotion processing outside attentional focus. This paradigm provides a new protocol for future studies on emotional prosody with potential extension from adults to infants, children, and people with difficulties in affective processing. Future work can also investigate the functional significance of MMN and P3a responses to emotional prosody in both auditory and visual modalities as

neural indices of predictive coding that may explain individual differences in subsequent behavioral judgment.

Chapter 5: General Discussion & Conclusions

Communication is inherently a parallel processing task that requires perceivers to undertake multi-layered information. Vocal emotional expression is one of the crucial speech information that helps listeners to fully understand the message. The current dissertation characterizes infant listeners' selective attention to different emotional prosodies and the relevant acoustic variables (Study 1). Infants' neurophysiological responses further reveal the developmental changes and early sex differences in the processing of emotional speech (Study 1 and Study 2). The development of neural activities and the emergence of sex/gender differences in emotional prosody processing can be further compared across infants and adults' EEG responses recorded by the same testing protocol (Study 2 and Study 3). As for experienced adult listeners, emotional prosodies are represented differently in male and female listeners' pre-attentive auditory processing (Study 3). The following sections will provide cross-study comparisons and discussions.

I. General Discussion

A. Developmental Changes in the Neural but not Behavioral Responses to Emotions

There are consistent findings as well as inconsistencies in infants' behavioral and EEG data (Study 1 and Study 2). Infants showed more listening attention to affective over the neutral prosody by engaging in the emotional sound presentation longer in the behavioral task. To behaviorally demonstrate this listening bias, infants' central auditory system should already be able to reliably distinguish the emotional voices from the

neutral tone. Indeed, infants' early and late MMRs to each vocal emotion confirm that their pre-attentive system supports the later voluntary responses to speech with emotions. Comparing across different emotions, infants behaviorally orient to happy and sad sounds more than angry sounds. A similar pattern is also observed in their early and late MMRs that show similar neural sensitivity development in listening to happy and sad voices compared to angry voices.

One major discrepancy between infants' attentive (behavioral) and pre-attentive (neurophysiological) responses to emotional prosody is the age effect. The developmental changes in emotional speech processing are only observed in infants' pre-attentive neural activities but not overt looking times. For instance, younger infants show distinct MMRs to happy and angry prosodies, but not the older infants; infants of both age groups attend to happy and angry voices differently in the behavioral task. Younger infants' higher neural sensitivities to these two high-arousal vocal emotions underlie the fact that prosodic information may be more relevant than the linguistic information for very young infants, for their limited knowledge of language-specific information (Walker-Andrews, 2008). Considering both attentive and pre-attentive data, older infants may be experienced in emotional speech and can react appropriately without their automatic auditory systems registering the emotion category.

B. The Developmental Changes of Neural Sensitivity to Emotional Speech

Discussing the functional significance of the maturation of MMRs in infancy is still controversial because of the huge across-infant variability and some within-infant inconsistency. The following comparison describes the observed developmental trend of

the auditory MMRs to the emotional prosody change in a diverse listening context. Younger infants showed different early MMR amplitudes to happy, angry, and sad voices (Study 2), which is a great improvement from the indistinguishable MMRs to emotional voices in newborns (Kostilainen et al., 2018). Unlike younger infants' distinct early MMRs that captured the acoustic differences across the three emotions, older infants showed similar early MMRs to happy and sad speech with a more positive MMR to the angry speech (Study 2). Adult listeners also showed comparable MMNs to happy and sad speech, with an even more negative MMN to the angry speech (Study 3). Both older infants and adults' automatic change detection systems treat happy and sad sounds similarly and trigger a differential pre-attentive response to the angry voice, whereas the younger infants' central auditory system registers the three emotions as distinct categories. This developmental trend can be summarized as infants' initial acoustic-driven early neural sensitivity that later develops into a category-driven response to emotional prosody change similar to adults' auditory emotional MMN.

Both infant and adult listeners showed similar fronto-centrally oriented scalp topography of the later discriminatory component (late MMR in infants, P3a in adults) to the emotional prosody change. However, infants' late MMR may not be functionally comparable to adults' P3a here. Previous reports on newborns' or infants' neural sensitivities to emotional speech mainly expect the MMR around this 350 ms window to later develop into an adult-like negative MMR with a shorter latency (e.g., Cheng et al., 2012; Zhang et al., 2014). Furthermore, older infants' scalp topography showed a central negativity (Figure 6, Study 2), which resembles the adult MMN more than the adult P3a.

Therefore, a more positive late MMR may not be explained as more mature or more adult-like, even if adult listeners showed a positive P3a at this time window (Study 3).

C. The Emergence of Sex/Gender Differences in Emotional Prosody Processing

Sex/Gender differences have been observed in previous reports on emotional voice perception in adolescents and adults (Lambrecht et al., 2014; Lausen & Schacht, 2018; Paulmann et al., 2008), but infant studies seldom include biological sex as a factor in their processing of emotional information. It is thought that both biological (Chaplin, 2015; Everhart et al., 2009) and social factors (Brody & Hall, 2008; Keshtiari & Kuhlmann, 2016) contribute to the differential processing of emotions in male and female listeners. Study 2 in the current dissertation did not observe a biological sex effect on infants' early automatic processing of vocal emotions (in the first 200 ms), but a generally more negative-going late MMR in female infants (after 300 ms). In Study 3, both male and female adults show similar amplitudes of MMN and P3a, indicating a generally similar neural sensitivity to affective voices. However, the neural representation of each emotional voice differs within each gender group. The current dissertation adds empirical evidence to the literature supporting the biological factor in gender-related emotional processing differences by showing that infants' emotional MMR differs across their biological sex. Furthermore, the gender differences emerge from emotion-general (infants) to emotion-specific (adults) as listeners gain more experiences in the emotional speech.

II. Limitations

The missing link between emotion and language development motivates the current dissertation. Nonetheless, the three studies in this dissertation project can only address listeners' different attentive and pre-attentive processing of emotional cues in natural human speech, not the potential effects of socio-emotional cues on language development (or vice versa). There were few reports dedicated to testing whether emotional prosody facilitates word learning (e.g., Bhullar, 2008; Singh et al., 2004). None of the studies compared infants' performances of multiple emotions against the neutral one. Without enough empirical studies that systematically examine infants' perception of multiple emotional speech, it may be premature to discuss and draw conclusions on its implications for language development at this point in time.

Acoustic variables are crucial in delivering auditory emotion, and Study 1 demonstrated that they partly account for infants' listening attention to emotional speech. However, the current EEG analysis uses the average ERP values of each emotion as the dependent variable, so the trial-by-trial acoustic variations within the same emotion were no longer available for the final statistical models. Recently, a more advanced single-trial EEG analysis was introduced, which can map the correlation between the acoustic variables of speech utterances (e.g., temporal envelope) and the neural responses (Horton et al., 2014). However, this analysis requires high-quality EEG data from listeners who stay still throughout the task and as many trials as possible to achieve reliable analysis. Therefore, applying single-trial EEG analysis may not be feasible for infant participants. Previous EEG studies successfully recorded MMR in

sleeping newborns. Even though the similarities and differences across awake and asleep MMR need further quantification, recording sleeping infants' emotional MMR may be the next step in obtaining high-quality neurophysiological data for single-trial EEG analysis.

Finally, the current dissertation project examined only infant and adult listeners and thus leaves age gaps to illustrate the full developmental trajectory of emotional prosody perception. Childhood and adolescence are important stages of socio-cognitive development. For instance, children show better emotional voice categorization and are less affected by pitch cues (Quam & Swingley, 2012). Furthermore, one study measured emotional prosody recognition in late childhood and adolescents along with the participants' salivary testosterone levels, and the authors only observed the gender effect in adolescents that may be attributed to higher testosterone levels in the male adolescents (Fujisawa & Shinohara, 2011). Whether or not the hormone level is underlying the gender differences in emotional processing, future studies on emotional speech will need to include children and adolescents to fully address the developmental effect and the emergence of sex effect.

III. Implications and Future Directions

The overarching goal of this dissertation is to understand emotional processing in the context of speech and language, with a long-term goal of characterizing the role of socio-emotional information in language development. Understanding the interplay between emotional processing, speech perception, and language acquisition can provide the fundamental knowledge to better address the challenges in populations who struggle

with both socio-emotional and speech information. After the effect of emotional information on language learning is defined, better intervention strategies or classroom supports can be proposed.

Simultaneous bilingual infants grow up exposed to two different language systems and encounter more variant speech inputs compared to their monolingual counterparts. Even so, both monolingual and bilingual infants reach similar language developmental milestones (Ramirez-Esparza & Garcia-Sierra, 2014). Socio-emotional inputs are crucial in early language acquisition, but there is currently no report on bilingual infants' emotional speech perception. Since bilingual infants need to remain sensitive to more linguistic and prosodic cues for different language systems, they may show listening attention and neural sensitivities to emotional speech that are different from monolingual infants. For instance, word-learning context with highly varying emotional tones may distract 6-month-old monolinguals from paying attention to the phonetic cues (Singh, 2008), because monolingual infants may only attend to one most distinguishable cue to learn new words. It will be of great interest to know if 6-month-old bilinguals also attend to one cue or perhaps multiple cues in similar word-learning contexts, given their listening experience with more speech variants.

Children with Autism Spectrum Disorder (ASD) tend to have a hard time using socio-emotional information in their daily communication (Kanner, 1943; Zhang et al., 2021). They are less sensitive to the vocal expressions of emotions than non-autistic children (e.g., Brooks et al., 2018; Wang & Tsao, 2015), even if they can sometimes correctly label the emotional prosody (e.g, Baker et al., 2010). Their discrepant

performances in naming and using the emotional prosody can be further investigated using the multi-feature paradigm with EEG recording. If smaller emotional MMN and P3a are measured, it indicates that autistic children show a less automatic affective voice detection before making any behavioral response. Effortful emotional voice processing may tax these children's cognitive resources, leading to difficulties in using emotional signals in daily conversation effectively.

Children with cochlear implants (CIs) also have difficulty in emotional prosody perception because the rich spectral and temporal information in acoustic speech is greatly degraded after being transmitted through the CI device (Jiam et al., 2017; Pak & Katz, 2019; Van De Velde et al., 2019). One study that implemented the central fixation paradigm found that infants with CIs showed better listening attention to speech with exaggerated prosody (Wang et al., 2017), but the relevant acoustic variables were not included. Future studies can further investigate how young listeners with CIs are directed to emotional and acoustic signals in the listening environment. With sufficient empirical data, the connection between early processing of socio-emotional information and early speech acquisition can be better established.

At this time of high parental stress due to the COVID-19 pandemic, research related to the impact of socio-emotional inputs on early speech and language acquisition may bring awareness of infants' listening environment—whether or not infants are actively attending to the speech, and whether or not the speech is directed to them. Caregivers' vocal emotional expressions are crucial in facilitating early communication and forming infant-caregiver attachment. Emotional expressions in other sensory

modalities such as visual and tactile channels are also important in early development and worth studying in future research.

IV. Conclusion

The current dissertation employed a central fixation paradigm to characterize infants' attention to emotional prosody in natural speech utterances and used a multi-feature oddball paradigm to reveal the age and sex effects. Infants before the age of one listen more to the affective over neutral voices, indicating the importance of socio-emotional signals in drawing young listeners' attention to the speech signals. The successful attempt to include acoustic variables in explaining infants' selective attention to vocal emotions also set the stage for future studies to consider relevant acoustic signals that entail the affect in speech. In the multi-feature oddball task, both infants and adults show pre-attentive neurophysiological responses to different emotional prosodies, indicating that the ability to extract emotional prosodic categories from a complex linguistic context is already available for infants before the age of one. The sex differences in infants and adults' automatic emotional sound processing at the cortical level provide an impetus for future studies to elucidate the biological mechanism at play and its functional significance in linguistic and socio-emotional development.

References

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology*, *12*(2), 169-177. https://doi.org/https://doi.org/10.1016/S0959-4388(02)00301-X

Aldridge, M. (1994). *Newborns' perception of emotion in voices*. International Conference on Infant Studies, Paris, France.

Alho, K., Sajaniemi, N. K., Niittyvuopio, T., Sainio, K. O., & Näätänen, R. (1990). ERPs to an auditory stimulus change in preterm and fullterm infants. In *psychophysiological brain research* (pp. 139-142). Tilburg University Press.

Alpert, M., Pouget, E. R., & Silva, R. R. (2001). Reflections of depression in acoustic measures of the patient's speech. *Journal of Affective Disorders*, *66*(1), 59-69.

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

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105-1138. https://doi.org/10.1111/j.1551-6709.2011.01181.x

Aslin, R. N. (2007). What's in a look? *Developmental Science, 10*(1), 48-53.

Bachorowski, J.-A., & Owren, M. J. (2008). Vocal expressions of emotion. *Handbook of Emotions*, *3*, 196-210.

Bahn, D., Vesker, M., Schwarzer, G., & Kauschke, C. (2021). A multimodal comparison of emotion categorization abilities in children with developmental language disorder. *Journal of Speech, Language, and Hearing Research*, 1-15. https://doi.org/10.1044/2020_JSLHR-20-00413

Baker, K. F., Montgomery, A. A., & Abramson, R. (2010). Brief report: perception and lateralization of spoken emotion by youths with high-functioning forms of autism. *Journal of Autism and Developmental Disorders*, *40*(1), 123-129.

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

Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in Cognitive Sciences*, *11*(8), 327-332.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577-660.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., Green, P., & Bolker, M. B. (2015). Package 'Lme4'. *Convergence*, *12*(1), 2.

Ben-David, B. M., Multani, N., Shakuf, V., Rudzicz, F., & van Lieshout, P. H. (2016). Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *Journal of Speech, Language, and Hearing Research*, *59*(1), 72-89.

Benders, T. (2013). Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior & Development*, *36*(4), 847-862. https://doi.org/10.1016/j.infbeh.2013.09.001

Berman, J. M., Chambers, C. G., & Graham, S. A. (2010). Preschoolers' appreciation of speaker vocal affect as a cue to referential intent. *Journal of Experimental Child Psychology*, *107*(2), 87-99.

Bhullar, N. (2008). Effects of facial and vocal emotion on word recognition in 11-to-13-month-old infants. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *69*(5-B), 3293.

Blasi, A., Mercure, E., Lloyd-Fox, S., Thomson, A., Brammer, M., Sauter, D., Deeley, Q., Barker, G. J., Renvall, V., & Deoni, S. (2011). Early specialization for voice and emotion processing in the infant brain. *Current Biology, 21*(14), 1220-1224.

Boersma, P., & Weenink, D. (2020). *Praat: Doing Phonetics by Computer [Computer Program]. Version 6.1.09*.

Brody, L. R., & Hall, J. A. (2008). Gender and emotion in context. *Handbook of emotions*, *3*, 395-408.

Brooks, P. J., Gaggi, N. L., & Ploog, B. O. (2018). Generalization of content and emotional prosody across speakers varying in gender in youth with autism spectrum disorder. *Research in Developmental Disabilities*, *83*, 57-68.

Burkhardt, F., & Sendlmeier, W. F. (2000). *Verification of Acoustical Correlates of Emotional Speech Using Formant-Synthesis*. ISCA Tutorial and Research Workshop (ITRW) on speech and emotion.

Carminati, M., Fiori-Duharcourt, N., & Isel, F. (2018). Neurophysiological differentiation between preattentive and attentive processing of emotional expressions on French vowels. *Biological Psychology*, *132*, 55-63.

Chaplin, T. M. (2015). Gender and emotion expression: A developmental contextual perspective. *Emotion Review*, *7*(1), 14-21. https://doi.org/10.1177/1754073914544408

Charpentier, J., Kovarski, K., Roux, S., Houy-Durand, E., Saby, A., Bonnet-Brilhault, F., Latinus, M., & Gomot, M. (2018). Brain mechanisms involved in angry prosody change detection in school-age children and adults, revealed by electrophysiology. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(4), 748-763.

Cheng, Y., Lee, S.-Y., Chen, H.-Y., Wang, P.-Y., & Decety, J. (2012). Voice and emotion processing in the human neonatal brain. *Journal of Cognitive*

Neuroscience, 24(6), 1411-1419.

Cheour-Luhtanen, M., Alho, K., Kujala, T., Sainio, K., Reinikainen, K., Renlund, M., Aaltonen, O., Eerola, O., & Näätänen, R. (1995). Mismatch negativity indicates vowel discrimination in newborns. *Hearing Research, 82*(1), 53-58.

Cheour, M. (2007). Development of mismatch negativity (MMN) during infancy. In M. de Haan (Ed.), *Infant EEG and event-related potentials* (pp. 171–198). Psychology Press.

Cheour, M., Kushnerenko, E., Ceponiene, R., Fellman, V., & Näätänen, R. (2002). Electric brain responses obtained from newborn infants to changes in duration in complex harmonic tones. *Developmental Neuropsychology, 22*(2), 471-479.

Cheour, M., Leppänen, P. H., & Kraus, N. (2000). Mismatch negativity (MMN) as a tool for investigating auditory discrimination and sensory memory in infants and children. *Clinical Neurophysiology, 111*(1), 4-16.

Chong, S., Werker, J. F., Russell, J. A., & Carroll, J. M. (2003). Three facial expressions mothers direct to their infants. *Infant and Child Development: An International Journal of Research and Practice*, *12*(3), 211-232. https://doi.org/10.1002/icd.286

Cohen, L. B., Atkinson, D. J., & Chaput, H. H. (2000). *Habit 2000: A New Program for Testing Infant Perception and Cognition [Computer Program]. Version 1*. In The University of Texas, Austin, TX.

Conboy, B. T., Brooks, R., Meltzoff, A. N., & Kuhl, P. K. (2015). Social interaction in infants' learning of second-language phonetics: An exploration of brain–behavior relations. *Developmental Neuropsychology*, *40*(4), 216-229. https://doi.org/10.1080/87565641.2015.1014487

Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of Learning and Memory, 92*(2), 225-234.

Cooper, R. P., & Aslin, R. N. (1994). Developmental differences in infant attention to the spectral properties of infant-directed speech. *Child Development*, *65*(6), 1663-1677. https://doi.org/10.1111/j.1467-8624.1994.tb00841.x

Corbeil, M., Trehub, S. E., & Peretz, I. (2013). Speech vs. singing: Infants choose happier sounds. *Frontiers in Psychology*, *4*, 372. https://doi.org/10.3389/fpsyg.2013.00372

Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. *Handbook of Emotions*, *2*(2), 91-115.

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, *52*(4), 521-536. http://dx.doi.org/10.1037/dev0000083

Csibra, G., Kushnerenko, E., & Grossmann, T. (2008). Electrophysiological methods in studying infant cognitive development. In C. A. Nelson & M. Luciana (Eds.),

Handbook of developmental cognitive neuroscience (pp. 247–262). MIT Press.

Cunningham, S., Weinel, J., & Picking, R. (2018). High-level analysis of audio features for identifying emotional valence in human singing. In *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion* (pp. 1-4). https://doi.org/10.1145/3243274.3243313

Damasio, A. R. (2006). *Descartes' Error*. Random House.

Dammers, J., Schiek, M., Boers, F., Silex, C., Zvyagintsev, M., Pietrzyk, U., & Mathiak, K. (2008). Integration of amplitude and phase statistics for complete artifact removal in independent components of neuromagnetic recordings. *IEEE Transactions on Biomedical Engineering, 55*(10), 2353-2362.

Davidson, R. J. (2000). Cognitive neuroscience needs affective neuroscience (and vice versa). *Brain and Cognition*, *42*(1), 89-92.

De Haan, M. (2002). Introduction to infant EEG and event-related potentials. *Infant EEG and Event-Related Potentials*, 39-76.

De Haan, M. (Ed.). (2013). *Infant EEG and Event-Related Potentials*. Psychology Press.

Dehaene-Lambertz, G. (2000). Cerebral specialization for speech and non-speech stimuli in infants. *Journal of Cognitive Neuroscience, 12*(3), 449-460.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9-21. https://doi.org/https://doi.org/10.1016/j.jneumeth.2003.10.009

Delorme, A., Makeig, S., & Sejnowski, T. (2001). *Automatic artifact rejection for EEG data using high-order statistics and independent component analysis*. Proceedings of the Third International ICA Conference.

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.

D'Entremont, B., & Muir, D. (1999). Infant responses to adult happy and sad vocal and facial expressions during face-to-face interactions. *Infant Behavior and Development, 22*(4), 527-539.

Donges, U.-S., Kersting, A., & Suslow, T. (2012). Women's greater ability to perceive happy facial emotion automatically: Gender differences in affective priming. *PloS One*, *7*(7), e41745. https://doi.org/10.1371/journal.pone.0041745

Dupuis, K., & Pichora-Fuller, M. K. (2010). *Toronto Emotional Speech Set (Tess)*. University of Toronto, Psychology Department. https://doi.org/10.5683/SP2/E8H2MF

Eckert, P. (1989). The whole woman: Sex and gender differences in variation. *Language Variation and Change*, *1*(3), 245-267.

Ekman, P. (1984). Expression and the nature of emotion. *Approaches to Emotion*, *3*(19), 344.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572-587.

Escera, C., Alho, K., Schröger, E., & Winkler, I. W. (2000). Involuntary attention and distractibility as evaluated with event-related brain potentials. *Audiology and Neurotology*, *5*(3-4), 151-166.

Escera, C., Alho, K., Winkler, I., & Näätänen, R. (1998). Neural mechanisms of involuntary attention to acoustic novelty and change. *Journal of Cognitive Neuroscience*, *10*(5), 590-604.

Escera, C., Yago, E., & Alho, K. (2001). Electrical responses reveal the temporal dynamics of brain events during involuntary attention switching. *European Journal of Neuroscience*, *14*(5), 877-883.

Everhart, D. E., Shipley, A., & Demaree, H. A. (2009). Perception of emotional Prosody: Establishing a link between sex-related differences, brain development, and sex hormones. *The Role of Prosody in Affective Speech*, *97*, 157.

Fan, Y.-T., Hsu, Y.-Y., & Cheng, Y. (2013). Sex matters: N-back modulates emotional mismatch negativity. *Neuroreport*, *24*(9), 457-463.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science, 146*(3644), 668-670.

Fauconnier, G., & Turner, M. (2008). The Origin of Language as a Product of the Evolution of Modern Cognition. *Origin and evolution of languages: Approaches, models, paradigms*. Equinox.

Feldman Barrett, L., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in Cognitive Sciences*, *11*(8), 327–332. https://doi.org/10.1016/j.tics.2007.06.003

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development*, 1497-1510.

Fernald, A. (1993). Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Development, 64*(3), 657-674.

Fernald, A., Kermanschachi, N., & Lees, D. (1984). The rhythms & sounds of soothing: maternal vestibular, tactile, & auditory stimulation and infant state. *Infant Behavior and Development*, *7*, 114.

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, *10*(3), 279-293. https://doi.org/10.1016/0163-6383(87)90017-8

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to

newborns. *Developmental Psychology*, *20*(1), 104. https://doi.org/10.1037/0012-1649.20.1.104

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477-501. https://doi.org/10.1017/S0305000900010679

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. *Developmental Psycholinguistics: On-Line Methods in Children's Language Processing*, *44*, 97.

Filippi, P., Ocklenburg, S., Bowling, D. L., Heege, L., Güntürkün, O., Newen, A., & de Boer, B. (2017). More than words (and faces): evidence for a stroop effect of prosody in emotion word processing. *Cognition and Emotion*, *31*(5), 879-891.

Fischer, A. H., & Manstead, A. S. (2008). Social functions of emotion. *Handbook of Emotions*, *3*, 456-468.

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. https://doi.org/10.1037/0012-1649.43.1.238

Flom, R., & Whiteley, M. O. (2014, Jan). The dynamics of intermodal matching: Seven- and 12-month-olds' intermodal matching of affect. *European Journal of Developmental Psychology, 11*(1), 111-119.

Forgas, J. P. (2008). Affect and cognition. *Perspectives on Psychological Science*, *3*(2), 94-101.

Forgas, J. P. (2012). *Affect in social thinking and behavior*. Psychology Press.

Fox, N. A., & Calkins, S. D. (2003). The development of self-control of emotion: intrinsic and extrinsic influences. *Motivation and Emotion*, *27*(1), 7-26.

Friederici, A. D., Friedrich, M., & Weber, C. (2002). Neural manifestation of cognitive and precognitive mismatch detection in early infancy. *Neuroreport, 13*(10), 1251-1254.

Friedman, D., Cycowicz, Y. M., & Gaeta, H. (2001). The novelty P3: An event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neuroscience & Biobehavioral Reviews*, *25*(4), 355-373. https://doi.org/https://doi.org/10.1016/S0149-7634(01)00019-7

Frijda, N. H. (2000). The psychologists' point of view. *Handbook of Emotions*, *2*, 59-74.

Frühholz, S., Ceravolo, L., & Grandjean, D. (2011). Specific brain networks during explicit and implicit decoding of emotional prosody. *Cerebral Cortex*, *22*(5), 1107-1117. https://doi.org/10.1093/cercor/bhr184

Fujisawa, T. X., & Shinohara, K. (2011). Sex differences in the recognition of emotional prosody in late childhood and adolescence. *The Journal of Physiological Sciences*, *61*(5), 429. https://doi.org/10.1007/s12576-011-0156-9

Garcia-Garcia, M., Domínguez-Borràs, J., SanMiguel, I., & Escera, C. (2008). Electrophysiological and behavioral evidence of gender differences in the modulation of distraction by the emotional context. *Biological Psychology*, *79*(3), 307-316.

García-Sierra, A., Ramírez-Esparza, N., Wig, N., & Robertson, D. (2021). Language learning as a function of infant directed speech (IDS) in Spanish: Testing neural commitment using the positive-MMR. *Brain and Language, 212*, 104890.

Garrido, M. I., Kilner, J. M., Stephan, K. E., & Friston, K. J. (2009). The mismatch negativity: a review of underlying mechanisms. *Clinical Neurophysiology, 120*(3), 453-463.

Gentili, R. J., Rietschel, J. C., Jaquess, K. J., Lo, L.-C., Prevost, C. M., Miller, M. W., Mohler, J. M., Oh, H., Tan, Y. Y., & Hatfield, B. D. (2014). *Brain biomarkers based assessment of cognitive workload in pilots under various task demands*. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby) talk to me: the social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science*, *24*(5), 339-344. https://doi.org/10.1177/0963721415595345

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, *14*(1), 23-45.

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychological Science*, *8*(3), 316-339.

Goydke, K. N., Altenmüller, E., Möller, J., & Münte, T. F. (2004). Changes in emotional tone and instrumental timbre are reflected by the mismatch negativity. *Cognitive Brain Research*, *21*(3), 351-359.

Grossmann, T. (2010). The development of emotion perception in face and voice during infancy. *Restorative Neurology and Neuroscience*, *28*(2), 219-236. https://doi.org/10.3233/RNN-2010-0499

Grossman, T. (2013). The early development of processing emotions in face and voice. In *Integrating face and voice in person perception* (pp. 95-116). Springer Science + Business Media; US.

Grossmann, T., Striano, T., & Friederici, A. D. (2005). Infants' electric brain responses to emotional prosody. *NeuroReport: For Rapid Communication of Neuroscience*

Research, 16(16), 1825-1828. https://doi.org/10.1097/01.wnr.0000185964.34336.b1

Grossmann, T., Striano, T., & Friederici, A. D. (2006). Crossmodal integration of emotional information from face and voice in the infant brain. *Developmental Science, 9*(3), 309-315. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-7687.2006.00494.x

Groves, P. M., & Thompson, R. F. (1970). Habituation: A dual-process theory. *Psychological Review, 77*(5), 419.

Haith, M. M. (1980). *Rules that babies look by: The organization of newborn visual activity*. Lawrence Erlbaum Associates.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin*, *85*(4), 845.

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice*, *21*(5), 531-540. https://doi.org/https://doi.org/10.1016/j.jvoice.2006.03.002

Hartkopf, J., Moser, J., Schleger, F., Preissl, H., & Keune, J. (2019). Changes in event-related brain responses and habituation during child development–A systematic literature review. *Clinical Neurophysiology, 130*(12), 2238-2254.

He, C., Hotson, L., & Trainor, L. J. (2007). Mismatch responses to pitch changes in early infancy. *Journal of Cognitive Neuroscience*, *19*(5), 878-892.

He, C., Hotson, L., & Trainor, L. J. (2009). Maturation of cortical mismatch responses to occasional pitch change in early infancy: Effects of presentation rate and magnitude of change. *Neuropsychologia, 47*(1), 218-229.

Heider, F. (2013). *The psychology of interpersonal relations*. Psychology Press.

Hilgard, E. R. (1980). The trilogy of mind: Cognition, affection, and conation. *Journal of the History of the Behavioral Sciences*, *16*(2), 107-117.

Hirsh-Pasek, K., & Golinkoff, R. M. (1996). The intermodal preferential looking paradigm: A window onto emerging language comprehension. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for assessing children's syntax* (pp. 105–124). The MIT Press.

Hoemann, K., Xu, F., & Feldman Barrett, L. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental Psychology*, 55(9), 1830–1849. https://doi.org/10.1037/dev0000686

Hohenberger, A. (2011). The role of affect and emotion in language development. In *Affective computing and interaction: Psychological, cognitive and neuroscientific perspectives* (pp. 208-243). IGI Global.

Horton, C., Srinivasan, R., & D'Zmura, M. (2014). Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party'. *Journal of Neural Engineering*, *11*(4), 046015.

Horváth, J., Czigler, I., Jacobsen, T., Maess, B., Schröger, E., & Winkler, I. (2008). MMN or No MMN: No magnitude of deviance effect on the MMN amplitude. *Psychophysiology*, *45*(1), 60-69.

Houston, D. M. (1999). *The role of talker variability in infant word representations*. Unpublished doctoral dissertation. Johns Hopkins University, Baltimore.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582. https://doi.org/10.1037/0096-1523.26.5.1570

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development: An International Journal of Research and Practice*, *13*(4), 341-348.

Hung, A.-Y., & Cheng, Y. (2014). Sex differences in preattentive perception of emotional voices and acoustic attributes. *Neuroreport, 25*(7), 464-469.

Isen, A. M. (1987). Positive affect, cognitive processes, and social behavior. *Advances in Experimental Social Psychology*, *20*, 203-253.

Ishi, C., Ishiguro, H., & Hagita, N. (2010). Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*, 1-12. https://doi.org/10.1155/2010/528193

Jaywant, A., & Pell, M. D. (2012). Categorical processing of negative emotions from speech prosody. *Speech Communication*, *54*(1), 1-10. https://doi.org/10.1016/j.specom.2011.05.011

Jiam, N. T., Caldwell, M., Deroche, M. L., Chatterjee, M., & Limb, C. J. (2017). Voice emotion perception and production in cochlear implant users. *Hearing Research*, *352*, 30-39. https://doi.org/10.1016/j.heares.2017.01.006

Jiang, A., Yang, J., & Yang, Y. (2014). Mmn Responses During Implicit Processing of Changes in Emotional Prosody: An Erp Study Using Chinese Pseudo-Syllables. *Cognitive Neurodynamics*, *8*(6), 499-508.

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. *Handbook of Emotions*, *2*, 220-235.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child*, *2*(3), 217-250.

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.

Kiley Hamlin, J., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental Science*, *13*(6), 923-929.

Kim, S. K., & Sumner, M. (2017). Beyond lexical meaning: The effect of emotional prosody on spoken word recognition. *The Journal of the Acoustical Society of America*, *142*(1), EL49-EL55.

Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy*, *4*(1), 85-110. https://doi.org/10.1207/S15327078IN0401_5

Kitamura, C., & Notley, A. (2009). The shift in infant preferences for vowel duration and pitch contour between 6 and 10 months of age. *Developmental Science*, *12*(5), 706-714. https://doi.org/10.1111/j.1467-7687.2009.00818.x

Koerner, T. K., & Zhang, Y. (2015). Effects of background noise on inter-trial phase coherence and auditory N1–P2 responses to speech stimuli. *Hearing Research*, *328*, 113-119.

Kok, T. B., Post, W. J., Tucha, O., de Bont, E. S., Kamps, W. A., & Kingma, A. (2014). Social competence in children with brain disorders: A meta-analytic review. *Neuropsychology Review, 24*(2), 219-235.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, *15*(2), 99-117.

Korpilahti, P., Jansson-Verkasalo, E., Mattila, M.-L., Kuusikko, S., Suominen, K., Rytky, S., Pauls, D. L., & Moilanen, I. (2007). Processing of affective speech prosody is impaired in Asperger syndrome. *Journal of Autism and Developmental Disorders, 37*(8), 1539-1549.

Kostilainen, K., Partanen, E., Mikkola, K., Wikström, V., Pakarinen, S., Fellman, V., & Huotilainen, M. (2020). Neural processing of changes in phonetic and emotional speech sounds and tones in preterm infants at term age. *International Journal of Psychophysiology*, *148*, 111-118.

Kostilainen, K., Wikström, V., Pakarinen, S., Videman, M., Karlsson, L., Keskinen, M., Scheinin, N. M., Karlsson, H., & Huotilainen, M. (2018). Healthy full-term infants' brain responses to emotionally and linguistically relevant sounds using a multi-feature mismatch negativity (MMN) paradigm. *Neuroscience Letters*.

Kret, M. E., & De Gelder, B. (2012). A review on sex differences in processing emotional signals. *Neuropsychologia*, *50*(7), 1211-1221. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2011.12.022

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, *10*(1), 110-120. https://doi.org/10.1111/j.1467-7687.2007.00572.x

Kushnerenko, E., Ceponiene, R., Balan, P., Fellman, V., & Näätänen, R. (2002). Maturation of the auditory change detection response in infants: A longitudinal ERP study. *Neuroreport*, *13*(15), 1843-1848.

Kushnerenko, E. V., Van den Bergh, B. R., & Winkler, I. (2013). Separating acoustic deviance from novelty during the first year of life: a review of event-related potential evidence. *Frontiers in Psychology, 4*, 595.

Kushnerenko, E., Winkler, I., Horváth, J., Näätänen, R., Pavlov, I., Fellman, V., & Huotilainen, M. (2007). Processing acoustic change and novelty in newborn infants. *European Journal of Neuroscience*, *26*(1), 265-274.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Ladd, D. R., Silverman, K. E., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *The Journal of the Acoustical Society of America*, *78*(2), 435-444.

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.

Lausen, A., & Schacht, A. (2018). Gender differences in the recognition of vocal emotions. *Frontiers in Psychology*, *9*, 882.

Lawrence, L., & Fernald, A. (1993). *When prosody and semantics conflict: Infants' sensitivity to discrepancies between tone of voice and verbal content*. Poster presented at the Biennial Meeting of the Society for Research in Child Development.

Leibold, L. J., & Werner, L. A. (2007). Infant auditory sensitivity to pure tones and frequency-modulated tones. *Infancy*, *12*(2), 225-233. https://doi.org/10.1111/j.1532-7078.2007.tb00241.x

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R Package Version*, *1*(1), 3.

Leppänen, P. H., Guttorm, T. K., Pihko, E., Takkinen, S., Eklund, K. M., & Lyytinen, H. (2004). Maturational effects on newborn ERPs measured in the mismatch negativity paradigm. *Experimental Neurology, 190*, 91-101.

Leppänen, J. M., Moulson, M. C., Vogel-Farley, V. K., & Nelson, C. A. (2007). An ERP study of emotional face processing in the adult and infant brain. *Child Development*, *78*(1), 232–245. https://doi.org/10.1111/j.1467-8624.2007.00994.x

Leppänen, P. H., Pihko, E., Eklund, K. M., & Lyytinen, H. (1999). Cortical responses of infants with and without a genetic risk for dyslexia: II. Group effects. *Neuroreport*, *10*(5), 969-973.

Lin, Y., Ding, H., & Zhang, Y. (2020). Prosody dominates over semantics in emotion word processing: evidence from cross-channel and cross-modal stroop effects. *Journal of Speech, Language, and Hearing Research*, *63*(3), 896-912.

Lin, Y., Ding, H., & Zhang, Y. (In press). Gender differences in identifying facial, prosodic, and semantic emotions processing show category- and channel-specific effects mediated by encoder gender. *Journal of Speech, Language and Hearing Research*.

Liu, P., & Pell, M. D. (2012). Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli. *Behavior Research Methods*, *44*(4), 1042-1051. https://doi.org/10.3758/s13428-012-0203-3

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience, 8*, 213.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Makarova, V., & Petrushin, V. A. (2002). *Ruslana: A database of Russian emotional utterances*. Seventh International Conference on Spoken Language Processing.

Mani, N., & Pätzold, W. (2016). Sixteen-month-old infants' segment words from infant- and adult-directed speech. *Language Learning and Development*, *12*(4), 499-508. https://doi.org/10.1080/15475441.2016.1171717

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24-52. https://doi.org/10.1177/2515245919900809

Marty, A. (1908). Untersuchungen zur allgemeinen grundlegung der grammatik und sprachphilosophie. *Niemeyer, Halle/Saale*.

Masapollo, M., Polka, L., & Ménard, L. (2016). When infants talk, infants listen: Pre-babbling infants prefer listening to speech with infant vocal properties. *Developmental Science*, *19*(2), 318-328. https://doi.org/10.1111/desc.12298

Mastropieri, D., & Turkewitz, G. (1999). Prenatal experience and neonatal responsiveness to vocal expressions of emotion. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, *35*(3), 204-214. https://doi.org/10.1002/(SICI)1098-2302(199911)35:3%3C204::AID-DEV5%3E3.0.CO;2-V

Maurer, U., Bucher, K., Brem, S., & Brandeis, D. (2003). Development of the automatic mismatch response: from frontal positivity in kindergarten children to the mismatch negativity. *Clinical Neurophysiology, 114*(5), 808-817.

McCann, J., & Peppé, S. (2003). Prosody in autism spectrum disorders: a critical review. *International Journal of Language & Communication Disorders*, *38*(4), 325-350. https://doi.org/10.1080/1368282031000154204

McClure, E. B. (2000). A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychological Bulletin, 126*(3), 424.

Mehler, J., Bertoncini, J., Barriere, M., & Jassik-Gerschenfeld, D. (1978). Infant recognition of mother's voice. *Perception, 7*(5), 491-497.

Mehrabian, A., & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, *6*(1), 109.

Minagawa-Kawai, Y., Van Der Lely, H., Ramus, F., Sato, Y., Mazuka, R., & Dupoux, E. (2011). Optical brain imaging reveals general auditory and language-specific processing in early infant development. *Cerebral Cortex, 21*(2), 254-261.

Miyazawa, K., Shinya, T., Martin, A., Kikuchi, H., & Mazuka, R. (2017). Vowels in infant-directed speech: More breathy and more variable, but not clearer. *Cognition*, *166*, 84-93. https://doi.org/10.1016/j.cognition.2017.05.003

Mokhsin, M. B., Rosli, N. B., Adnan, W. A. W., & Manaf, N. A. (2014). Automatic music emotion classification using artificial neural network based on vocal and instrumental sound timbres. In *SoMeT* (pp. 3-14). https://doi.org/10.3233/978-1-61499-434-3-3

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development, 16*(4), 495-500.

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

Mumme, D. L., Fernald, A. & Herrera, C. (1996). Infants' responses to facial and vocal emotional signals in a social referencing paradigm. *Child Development*, *67*(6), 3219-3237. https://doi.org/10.1111/j.1467-8624.1996.tb01910.x

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*(2), 1097-1108. https://doi.org/10.1121/1.405558

Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, *42*(4), 313-329.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, *118*(12), 2544-2590.

Näätänen, R., Pakarinen, S., Rinne, T., & Takegata, R. (2004). The mismatch negativity (MMN): Towards the optimal paradigm. *Clinical Neurophysiology*, *115*(1), 140-144.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., & Winkler, I. (2001). 'Primitive intelligence' in the auditory cortex. *Trends in Neurosciences*, *24*(5),

283-288.

Nagy, E., Potts, G. F., & Loveland, K. A. (2003). Sex-related ERP differences in deviance detection. *International Journal of Psychophysiology*, *48*(3), 285-292.

Neisser, U., & Hyman, I. (2000). *Memory observed: Remembering in natural contexts*. Macmillan.

Nelson, C. A., & De Haan, M. (1996). Neural correlates of infants' visual responsiveness to facial expressions of emotion. *Developmental Psychobiology*, *29*(7), 577–595. https://doi.org/10.1002/(SICI)1098-2302(199611)29:7<577::AID-DEV3>3.0.CO;2-R

Oakes, L. M. (2010). Using habituation of looking time to assess mental processes in infancy. *Journal of Cognition and Development, 11*(3), 255-268.

O'connor, K. (2012). Auditory processing in autism spectrum disorder: A review. *Neuroscience & Biobehavioral Reviews*, *36*(2), 836-854. https://doi.org/10.1016/j.neubiorev.2011.11.008

Otte, R. A., Donkers, F. C. L., Braeken, M. A. K. A., & Van den Bergh, B. R. H. (2015). Multimodal processing of emotional information in 9-month-old infants I: Emotional faces and voices. *Brain and Cognition, 95*, 99-106. https://doi.org/https://doi.org/10.1016/j.bandc.2014.09.007

Pak, C. L., & Katz, W. F. (2019). Recognition of emotional prosody by Mandarin-speaking adults with cochlear implants. *The Journal of the Acoustical Society of America*, *146*(2), EL165-EL171.

Pakarinen, S., Lovio, R., Huotilainen, M., Alku, P., Näätänen, R., & Kujala, T. (2009). Fast multi-feature paradigm for recording several mismatch negativities (MMNs) to phonetic and acoustic changes in speech sounds. *Biological Psychology*, *82*(3), 219-226.

Pakarinen, S., Sokka, L., Leinikka, M., Henelius, A., Korpela, J., & Huotilainen, M. (2014). Fast determination of MMN and P3a responses to linguistically and emotionally relevant changes in pseudoword stimuli. *Neuroscience Letters*, *577*, 28-33.

Pakarinen, S., Takegata, R., Rinne, T., Huotilainen, M., & Näätänen, R. (2007). Measurement of extensive auditory discrimination profiles using the mismatch negativity (MMN) of the auditory event-related potential (ERP). *Clinical Neurophysiology*, *118*(1), 177-185.

Panneton, R., Kitamura, C., Mattock, K., & Burnham, D. (2006). Slow speech enhances younger but not older infants' perception of vocal emotion. *Research in Human Development*, *3*(1), 7-19. https://doi.org/10.1207/s15427617rhd0301_2

Paris, M., Mahajan, Y., Kim, J., & Meade, T. (2018). Emotional speech processing deficits in bipolar disorder: The role of mismatch negativity and P3a. *Journal of Affective Disorders*, *234*, 261-269.

Paquette-Smith, M., & Johnson, E. K. (2016). I don't like the tone of your voice: Infants use vocal affect to socially evaluate others. *Infancy*, *21*(1), 104-121. https://doi.org/10.1111/infa.12098

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

Paulmann, S., & Kotz, S. A. (2008). Early emotional prosody perception based on different speaker voices. *Neuroreport*, *19*(2).

Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). Functional contributions of the basal ganglia to emotional prosody: Evidence from ERPs. *Brain Research*, *1217*, 171-178.

Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). How aging affects the recognition of emotional speech. *Brain and Language*, *104*(3), 262-269.

Perlovsky, L. (2009). Language and cognition. *Neural Networks*, *22*(3), 247-257.

Peter, V., Kalashnikova, M., Santos, A., & Burnham, D. (2016). Mature neural responses to infant-directed speech but not adult-directed speech in pre-verbal infants. *Scientific Reports*, *6*(1), 1-14.

Picton, T. W., Alain, C., Otten, L., Ritter, W., & Achim, A. (2000). Mismatch negativity: Different water in the same river. *Audiology and Neurotology*, *5*(3-4), 111-139.

Pinheiro, A. P., Barros, C., Dias, M., & Kotz, S. A. (2017). Laughter catches attention! *Biological Psychology*, *130*, 11-21. https://doi.org/https://doi.org/10.1016/j.biopsycho.2017.09.012

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128-2148.

Quam, C., & Swingley, D. (2012). Development in children's interpretation of pitch cues to emotions. *Child Development*, *83*(1), 236-250.

Rakison, D. H., & Yermolayeva, Y. (2010). Infant categorization. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 894-905.

Ramírez-Esparza, N., & García-Sierra, A. (2014). The bilingual brain: language, culture and identity. *The Oxford handbook of multicultural identity*, 35-56.

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*(6), 880-891. https://doi.org/10.1111/desc.12172

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145.

Salisch, M. v. (2001). Children's emotional development: Challenges in their relationships to parents, peers, and friends. *International Journal of Behavioral Development*, *25*(4), 310-319.

SanMiguel, I., Morgan, H. M., Klein, C., Linden, D., & Escera, C. (2010). On the functional significance of novelty-P3: Facilitation by unexpected novel sounds. *Biological Psychology*, *83*(2), 143-152. https://doi.org/https://doi.org/10.1016/j.biopsycho.2009.11.012

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. *Approaches to Emotion*, *2293*(317), 31.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*(2), 143.

Scherer, K. R. (1989). Vocal correlates of emotional arousal and affective disturbance. In *Handbook of social psychophysiology.* (pp. 165-197). John Wiley & Sons.

Schirmer, A., & Escoffier, N. (2010). Emotional MMN: Anxiety and heart rate correlate with the ERP signature for auditory change detection. *Clinical Neurophysiology*, *121*(1), 53-59.

Schirmer, A., & Kotz, S. A. (2003). ERP evidence for a sex-specific stroop effect in emotional speech. *Journal of Cognitive Neuroscience*, *15*(8), 1135-1148.

Schirmer, A., Kotz, S. A., & Friederici, A. D. (2002). Sex differentiates the role of emotional prosody during word processing. *Cognitive Brain Research*, *14*(2), 228-233. https://doi.org/10.1016/S0926-6410(02)00108-8

Schirmer, A., Kotz, S. A., & Friederici, A. D. (2005). On the role of attention for the processing of emotions in speech: Sex differences revisited. *Cognitive Brain Research*, *24*(3), 442-452.

Schirmer, A., Striano, T., & Friederici, A. D. (2005). Sex differences in the preattentive processing of vocal emotional expressions. *Neuroreport*, *16*(6), 635-639.

Schmid, P. C., Mast, M. S., Bombari, D., & Mast, F. W. (2011). Gender effects in information processing on a nonverbal decoding task. *Sex Roles*, *65*(1-2), 102-107.

Schröder, M. (2001). *Emotional speech synthesis: A review*. Seventh European Conference on Speech Communication and Technology.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). *Acoustic correlates of emotion dimensions in view of speech synthesis*. Seventh European Conference on Speech Communication and Technology.

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

Shafer, V. L., Yan, H. Y., & Garrido-Nag, K. (2012). Neural mismatch indices of vowel

discrimination in monolingually and bilingually exposed infants: Does attention matter? *Neuroscience Letters, 526*(1), 10-14.

Shouse, E. (2005). Feeling, emotion, affect. *M/c Journal*, *8*(6).

Shultz, S., & Vouloumanos, A. (2010). Three-month-olds prefer speech to other naturally occurring signals. *Language Learning and Development*, *6*(4), 241-257. https://doi.org/10.1080/15475440903507830

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, *106*(2), 833-870. https://doi.org/10.1016/j.cognition.2007.05.002

Singh, L., Morgan, J. L., & Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? *Infancy*, *3*(3), 365-394. https://doi.org/10.1207/S15327078IN0303_5

Singh, L., Morgan, J. L., & White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language*, *51*(2), 173-189. https://doi.org/10.1016/j.jml.2004.04.004

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501-532. https://doi.org/10.1016/j.dr.2007.06.002

Soken, N. H., & Pick, A. D. (1992). Intermodal perception of happy and angry expressive behaviors by seven-month-old infants. *Child Development*, *63*(4), 787-795. https://doi.org/10.1111/j.1467-8624.1992.tb01661.x

Soken, N. H., & Pick, A. D. (1999). Infants' perception of dynamic affective expressions: Do infants distinguish specific expressions? *Child Development*, *70*(6), 1275-1282. https://doi.org/10.1111/1467-8624.00093

Stager, C., & Werker, J. (1998). Methodological issues in studying the link between speech-perception and word learning. *Advances in Infancy Research*, *12*, 237-256.

Stern, D. N., Spieker, S., & MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Developmental Psychology*, *18*(5), 727. https://doi.org/10.1037/0012-1649.18.5.727

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: A review. *International Journal of Speech Technology*, *21*(1), 93-120.

Tato, R., Santos, R., Kompe, R., & Pardo, J. M. (2002). *Emotional space improves emotion recognition*. Seventh International Conference on Spoken Language Processing, Denver, CO.

Thierry, G., & Roberts, M. V. (2007). Event-related potential study of attention capture by affective sounds. *Neuroreport*, *18*(3), 245-248.

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.

Thönnessen, H., Boers, F., Dammers, J., Chen, Y.-H., Norra, C., & Mathiak, K. (2010). Early sensory encoding of affective prosody: Neuromagnetic tomography of emotional category changes. *Neuroimage*, *50*(1), 250-259.

Tillman, T. W., & Carhart, R. (1966). *An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University Auditory Test No. 6*.

Trainor, L. J. (2010). Using electroencephalography (EEG) to measure maturation of auditory cortex in infants: Processing pitch, duration and sound location. In Tremblay, R. E., Barr, R. G., Peters, R. de V., Boivin, M. (Eds.), *Encyclopedia on early childhood development.* Centre of excellence for early childhood development, Montreal, Quebec, 1-5.

Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science*, *11*(3), 188-195. https://doi.org/10.1111/1467-9280.00240

Trainor, L. J., Samuel, S. S., Desjardins, R. N., & Sonnadara, R. R. (2001). Measuring temporal resolution in infants using mismatch negativity. *Neuroreport*, *12*(11), 2443-2448.

Trainor, L. J., & Zacharias, C. A. (1998). Infants prefer higher-pitched singing. *Infant Behavior and Development*, *21*(4), 799-805. https://doi.org/10.1016/S0163-6383(98)90047-9

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: the negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383.

Vaish, A., & Striano, T. (2004). Is visual reference necessary? Contributions of facial versus vocal cues in 12-month-olds' social referencing behavior. *Developmental Science*, *7*(3), 261-269. https://doi.org/10.1111/j.1467-7687.2004.00344.x

Van De Velde, D. J., Schiller, N. O., Levelt, C. C., Van Heuven, V. J., Beers, M., Briaire, J. J., & Frijns, J. H. (2019). Prosody perception and production by children with cochlear implants. *Journal of Child Language*, *46*(1), 111-141.

Walker, A. S. (1982). Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology*, *33*, 514-535. https://doi.org/10.1016/0022-0965(82)90063-7

Walker-Andrews, A. S. (1986). Intermodal perception of expressive behaviors: Relation of eye and voice? *Developmental Psychology*, *22*(3), 373. https://doi.org/10.1037/0012-1649.22.3.373

Walker-Andrews, A. S. (2008). Intermodal emotional processes in infancy. *Handbook of Emotions*, 364-375.

Walker-Andrews, A. S., & Grolnick, W. (1983). Discrimination of vocal expressions by

young infants. *Infant Behavior & Development*. https://doi.org/10.1016/S0163-6383(83)90331-4

Walker-Andrews, A. S., & Lennon, E. (1991). Infants' discrimination of vocal expressions: Contributions of auditory and visual information. *Infant Behavior and Development*, *14*(2), 131-142. https://doi.org/10.1016/0163-6383(91)90001-9

Walker, A. S. (1982). Intermodal perception of expressive behaviors by human infants. *Journal of Experimental Child Psychology, 33*(3), 514-535.

Wambacq, I. J., & Jerger, J. F. (2004). Processing of affective prosody and lexical-semantics in spoken utterances as differentiated by event-related potentials. *Cognitive Brain Research*, *20*(3), 427-437.

Wang, J.-E., & Tsao, F.-M. (2015). Emotional prosody perception and its association with pragmatic language in school-aged children with high-function autism. *Research in Developmental Disabilities*, *37*, 162-170.

Wang, Y., Bergeson, T. R., & Houston, D. M. (2017). Infant-directed speech enhances attention to speech in deaf infants with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *60*(11), 3321-3333.

Wanrooij, K., Boersma, P., & Van Zuijen, T. (2014). Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study. *Frontiers in Psychology, 5*, 77.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*(2), 219.

Werker, J. F., & Fennell, C. T. (2009). Infant speech perception and later language acquisition: Methodological underpinnings. In J. Colombo, P. McCardle, & L. Freund (Eds.), *Infant pathways to language: Methods, models, and research disorders* (pp. 85–98). Psychology Press.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, *52*(4B), 1238-1250.

Williams, K. D. (2002). *Ostracism: The power of silence*. Guilford Press.

Winkler, I., Kushnerenko, E., Horváth, J., Čeponienė, R., Fellman, V., Huotilainen, M., Näätänen, R., & Sussman, E. (2003). Newborn infants can organize the auditory world. *Proceedings of the National Academy of Sciences, 100*(20), 11812-11815.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52-69.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151.

Zhang, Y., Koerner, T., Miller, S., Grice-Patil, Z., Svec, A., Akbari, D., ... & Carney, E.

(2011). Neural coding of formant-exaggerated speech in the infant brain. *Developmental Science*, *14*(3), 566-581.

Zhang, D., Liu, Y., Hou, X., Sun, G., Cheng, Y., & Luo, Y. (2014). Discrimination of fearful and angry emotional voices in sleeping human neonates: A study of the mismatch brain responses. *Frontiers in Behavioral Neuroscience, 8*, 422.

Zhang, M., Xu, S., Chen, Y., Lin, Y., Ding, H., & Zhang, Y. (2021). Recognition of affective prosody in autism spectrum conditions: A systematic review and meta-analysis. *Autism*. https://doi.org/10.1177/1362361321995725

Zora, H., Rudner, M., & Montell Magnusson, A. K. (2020). Concurrent affective and linguistic prosody with the same emotional valence elicits a late positive ERP response. *European Journal of Neuroscience*, *51*(11), 2236-2249.

Appendix A

The Four Wordlists

List 1	List 2	List 3	List 4
Merge	Shirt	Chat	Void
Germ	Which	Chair	Mill
Yes	Tool	Bar	Chair
Base	Germ	Tool	Base
Match	Sail	Dog	Merge
Chat	Turn	Which	Sail
Mill	Choice	Shirt	Shack
Bar	Dog	Choice	Yes
Void	Shack	Match	Turn
Shack	Chair	Yes	Shirt
Tool	Match	Sail	Bar
Which	Base	Merge	Dog
Turn	Merge	Void	Germ
Chair	Bar	Mill	Which
Shirt	Void	Germ	Choice
Dog	Chat	Base	Match
Sail	Mill	Turn	Chat
Choice	Yes	Shack	Tool

Appendix B

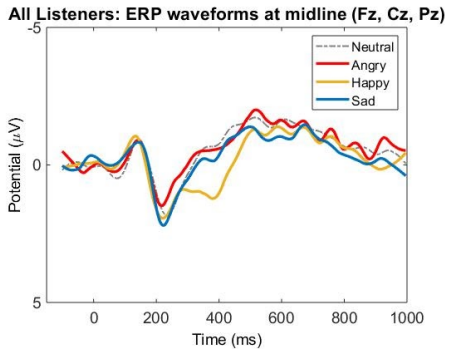
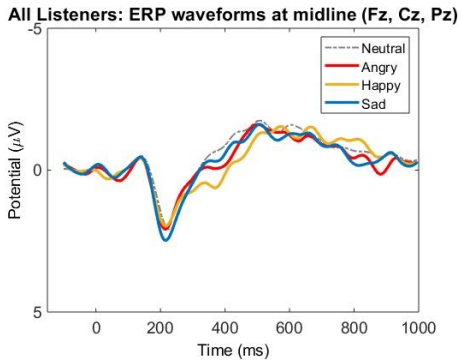
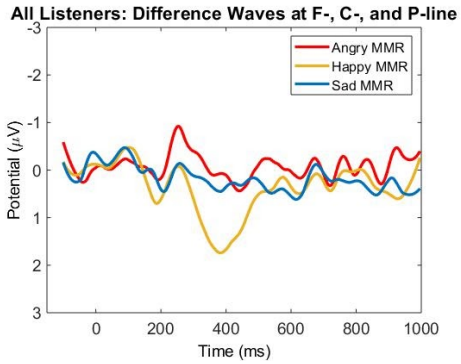
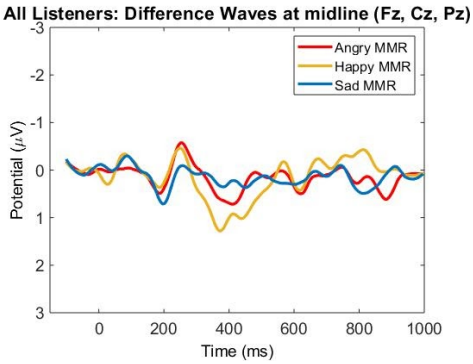
Two hundred phonetically balanced monosyllabic words from Northwestern University Auditory Test No. 6, NU-6 (Tillman & Carhart, 1966)

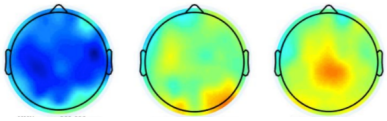
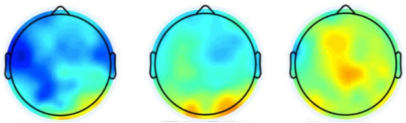
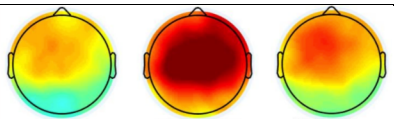
Back	Check	Fat	Hire	Late	Met	Peg	Rose	South	Vine
Bar	Cheek	Fit	Hit	Laud	Mill	Perch	Rot	Sub	Voice
Base	Chief	Five	Hole	Lean	Mob	Phone	Rough	Such	Void
Bath	Choice	Food	Home	Learn	Mode	Pick	Rush	Sure	Vote
Bean	Cool	Gap	Hurl	Lease	Mood	Pike	Said	Take	Wag
Beg	Dab	Gas	Hush	Lid	Moon	Pole	Sail	Talk	Walk
Bite	Date	Gaze	Jail	Life	Mop	Pool	Search	Tape	Wash
Boat	Dead	Germ	Jar	Limb	Mouse	Puff	Seize	Team	Week
Bone	Death	Get	Join	Live	Nag	Rag	Sell	Tell	Wheat
Book	Deep	Gin	Judge	Loaf	Name	Raid	Shack	Thin	When
Bought	Dime	Goal	Jug	Long	Near	Rain	Shall	Third	Which
Burn	Dip	Good	Juice	Lore	Neat	Raise	Shawl	Thought	Whip
Cab	Ditch	Goose	Keen	Lose	Nice	Rat	Sheep	Thumb	White
Calm	Dodge	Gun	Keep	Lot	Note	Reach	Shirt	Time	Wife
Came	Dog	Half	Keg	Love	Numb	Read	Should	Tip	Wire
Cause	Doll	Hall	Kick	Luck	Pad	Red	Shout	Tire	Witch
Chain	Door	Hash	Kill	Make	Page	Ring	Size	Ton	Yearn
Chair	Fail	Hate	King	Match	Pain	Ripe	Soap	Tool	Yes
Chalk	Fall	Have	Kite	Merge	Pass	Road	Soup	Tough	Young
Chat	Far	Haze	Knock	Mess	Pearl	Room	Sour	Turn	Youth

Appendix C

Comparisons of ERP waveforms, topographies, and statistical analyses—epochs extracted relative to sound onsets and vowel onsets.

Standard ERP analysis uses the onset of an auditory event to mark the beginning of an epoch. Even though emotional prosody is largely characterized by the fundamental frequency and intensity, which may be expressed more at vowels, consonants also carry clear emotional messages (e.g., affricate in angry voices). Appendix C presents side-by-side comparisons of ERP waveforms, topographies, and statistical analyses to show the similarity between epochs extracted relative to the sound onset (i.e., consonant onset) and vowel onset.

	Sound onset	**Vowel onset**
ERP waveforms		
Difference waveforms		

MMN topography		
P3a topography		
MMN statistical analysis	*Trial-level fixed factors* Emotion $F(2,572) = 8.21$*** Region $F(2, 572) = 1.27$ Laterality $F(2,572) = 0.59$ *Participant-level fixed factor* Gender $F(1,22) = 0.001$ *Cross-level interaction* Emotion*Gender $F(2,572) = 4.68$**	*Trial-level fixed factors* Emotion $F(2,572) = 6.11$** Region $F(2, 572) = 0.81$ Laterality $F(2,572) = 0.24$ *Participant-level fixed factor* Gender $F(1,22) = 0.02$ *Cross-level interaction* Emotion*Gender $F(2,572) = 1.39$
P3a statistical analysis	*Trial-level fixed factors* Emotion $F(2,572) = 25.67$*** Region $F(2, 572) = 26.13$*** Laterality $F(2,572) = 1.54$ *Participant-level fixed factor* Gender $F(1,22) = 0.49$ *Cross-level interaction* Emotion*Gender $F(2,572) = 14.61$**	*Trial-level fixed factors* Emotion $F(2,572) = 38.48$*** Region $F(2, 572) = 16.8$*** Laterality $F(2,572) = 0.93$ *Participant-level fixed factor* Gender $F(1,22) = 1.4$ *Cross-level interaction* Emotion*Gender $F(2,572) = 6.4$**