ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

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

Chieh Kao,[a] Maria D. Sera,[b] and Yang Zhang[a,c]

[a] Department of Speech-Language-Hearing Sciences, University of Minnesota, Twin Cities, Minneapolis [b] Institute of Child Development, University of Minnesota, Twin Cities, Minneapolis [c] Center for Neurobehavioral Development, University of Minnesota, Twin Cities, Minneapolis

ARTICLE INFO

Article History:
Received April 25, 2021
Revision received July 18, 2021
Accepted October 7, 2021

Editor-in-Chief: Bharath Chandrasekaran
Editor: Chao-Yang Lee

https://doi.org/10.1044/2021_JSLHR-21-00234

ABSTRACT

Purpose: The aim of this study was to investigate infants' listening preference for emotional prosodies in spoken words and identify their acoustic correlates.
Method: Forty-six 3- to-12-month-old infants (M_{age} = 7.6 months) completed a central fixation (or look-to-listen) paradigm in which four emotional prosodies (happy, sad, angry, and neutral) were presented. Infants' looking time to the string of words was recorded as a proxy of their listening attention. Five acoustic variables—mean fundamental frequency (F0), word duration, intensity variation, harmonics-to-noise ratio (HNR), and spectral centroid—were also analyzed to account for infants' attentiveness to each emotion.
Results: Infants generally preferred affective over neutral prosody, with more listening attention to the happy and sad voices. Happy sounds with breathy voice quality (low HNR) and less brightness (low spectral centroid) maintained infants' attention more. Sad speech with shorter word duration (i.e., faster speech rate), less breathiness, and more brightness gained infants' attention more than happy speech did. Infants listened less to angry than to happy and sad prosodies, and none of the acoustic variables were associated with infants' listening interests in angry voices. Neutral words with a lower F0 attracted infants' attention more than those with a higher F0. Neither age nor sex effects were observed.
Conclusions: This study provides evidence for infants' sensitivity to the prosodic patterns for the basic emotion categories in spoken words and how the acoustic properties of emotional speech may guide their attention. The results point to the need to study the interplay between early socioaffective and language development.

Language development takes place in a socioemotional 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' socioemotional and communicative skills. However, 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 (Barrett et al., 2007; 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

Correspondence to Yang Zhang: zhanglab@umn.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

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 nonwords 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 [IDS]) may encourage infants to extract the invariant phonetic features and promote a more robust word representation (Apfelbaum & 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 (F0) may be the secondary determinant. This study followed up this idea by including the angry prosody and examined within-infant listening preference to the 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 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).

Although 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 this 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 (a) mean F0, (b) word duration, (c) intensity variation, (d) HNR, and (e) 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). Although 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, 3-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, 5-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 5 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 1 year 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 prosodies. There is evidence for an early developmental change as infants younger than 5 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 (pairwise) 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 7 months, they can differentiate between happy and neutral sounds even when some asynchronous talking face videos were presented (Walker, 1982). By 9 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 to show 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 IDS (The 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 et al. (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 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 IDS—a speech style that is closely related to emotional speech. There is a consensus that infants prefer IDS to ADS (The 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 was 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 6 months attend more to words with longer duration, whereas infants older than 8 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 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.

This Study

This 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 F0, intensity variation, word duration, HNR, and spectral centroid—to examine their roles in infants' listening attention to emotional speech. These five acoustic parameters were all included as trial-level fixed factors to examine their roles in infants' listening attention to the four emotions. 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 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 this 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 et al. (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.

Other than age differences, our study also examined infants' sex differences in their 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.

Method

Participants

The final sample for statistical reports included 43 infants between the ages of 2 months 26 days and 11 months 11 days (male = 22, female = 21; M_{age} = 7.6 months or 231 days). Initially, 46 typically developing infants from 3 to 12 months (male = 25, female = 21; M_{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.

Materials

The speech stimuli included 18 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; Tillman & Carhart, 1966). The recordings of the words in different emotional prosodies were from the Toronto Emotional Speech Set (Dupuis & Pichora-Fuller, 2010). 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 F0, duration, intensity variation, 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.

We used a customized Praat script to concatenate the 18 words with the same emotional prosody into a 32-s trial, with 1-s silence between adjacent words. Four randomized word orders were created for word concatenation (see the Appendix 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 that were presented in a randomized block design. To familiarize the infants with the listening procedure, we also included a 32-s music clip with piano and theremin (an electronic musical instrument) as the pretest stimulus.

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 in. away from a 22-in. LCD monitor. A video camera was placed 8 in. 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.

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 s or when the 32-s 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-s 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 pseudorandomized. 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 min.

Table 1. The acoustic properties of each emotional prosody.

	Mean F0 (Hz)		Duration (ms)		Intensity variation (dB)		HNR (dB)		Spectral centroid (Hz)	
Emotions	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Angry	216.88	25.91	661	107	10.3	3.74	7.89	5.38	2160.65	1002.31
Happy	223.61	9.09	756	110	10.19	3.9	16.68	4.07	1151.06	264.22
Sad	174.58	25.44	831	106	9.59	3.51	17.5	3.75	630.87	405.11
Neutral	190.13	6.32	684	58	7.84	3.46	17.03	4.64	850.37	198.2

Note. The averaged values and standard deviations 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.

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, then the trials would be removed without any data interpolation or replacement (four trials were removed). The trials with listening times shorter than 1 s (10 trials) or reaching the maximum length of the sound file (one trial) were also excluded from further analysis (Shultz & Vouloumanos, 2010). All participants had two or more trials in 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 s, which corresponded to the middle of the 10th word in the original sound file, we averaged the mean F0, intensity variation, word duration, 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 predict 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 levels), 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[1]), trial number (1–16), mean F0 (numerical variable in Hertz), intensity variation (numeral variable in dB), word duration (numerical variable in second), 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 months). 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.[2]

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-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[3] 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, $F(3, 610) = 21.89$, $p < .001$; mean F0, $F(1, 622) = 81.65$, $p < .001$; word duration, $F(1, 528) = 31.82$, $p < .001$; intensity variation, $F(1, 625) = 4.96$, $p = .03$; and trial number, $F(1, 593) = 41.24$, $p < .001$, 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 (*p*s < .001); they listened longer to happy than angry voices ($p < .001$) and to sad than angry voices ($p = .003$). Infants listened more to words with lower mean F0, to words with shorter durations (i.e., faster speaking rate), and to words with greater intensity

[1] 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.

[2] 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, and emotion and sex in the final model (see the first paragraph in the Results 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)*

[3] $F(3, 630.5) = 0.35$, $p = .79$.

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.

Factor	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 × Mean F0	32.57***		32.94***
Emotion × Word Duration	9.47***		9.97***
Emotion × Intensity Variation	9.78***		10.00***
Emotion × HNR	35.00***		35.40***
Emotion × Spectral Centroid	30.24***		30.97***
Participant-level fixed factors			
Age	0.01	0.16	
Sex	0.67	1.17	
Cross-level interactions			
Age × Emotion	0.65		
Sex × Emotion	0.41		
Goodness-of-fit (deviance)	1006.7	1428.4	1009.4

*$p < .05$. ***$p < .001$.

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, $F(3, 620) = 34.08$, $p < .001$; word duration, $F(3, 624) = 11.73$, $p < .001$; intensity variation, $F(3, 630) = 14.52$, $p < .001$; HNR, $F(3, 630) = 38.36$, $p < .001$; and spectral centroid, $F(3, 624) = 32.84$, $p < .001$, were all significant. Figure 2

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).

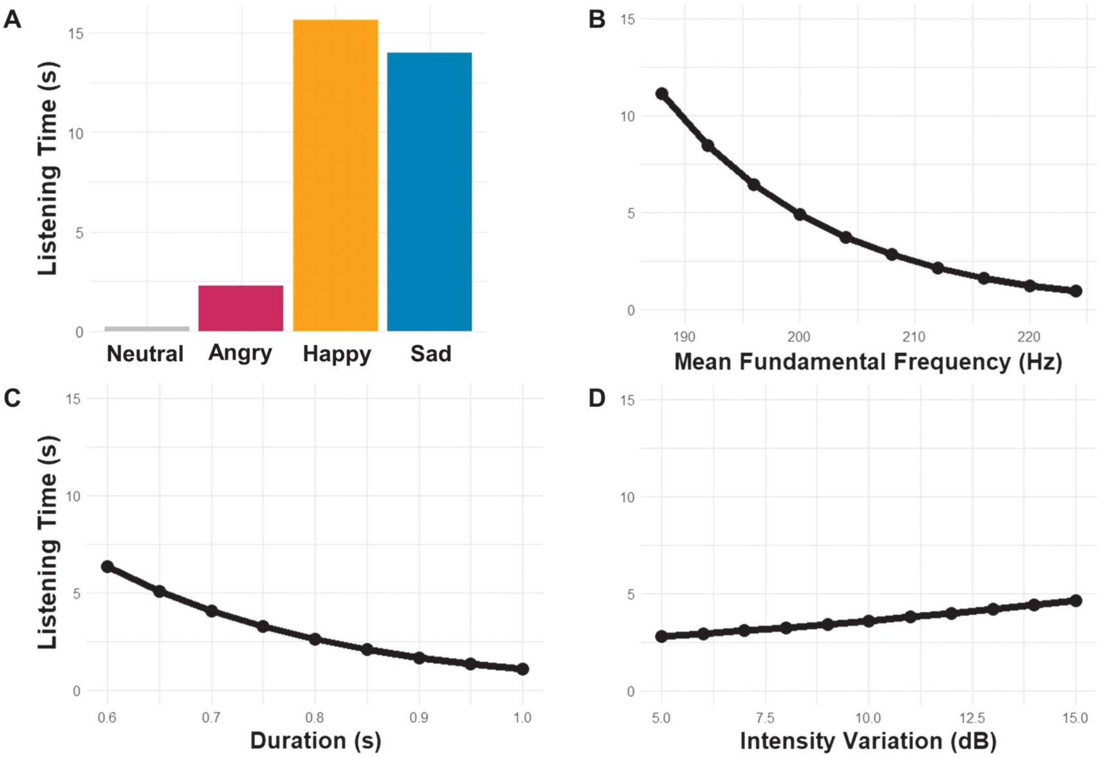
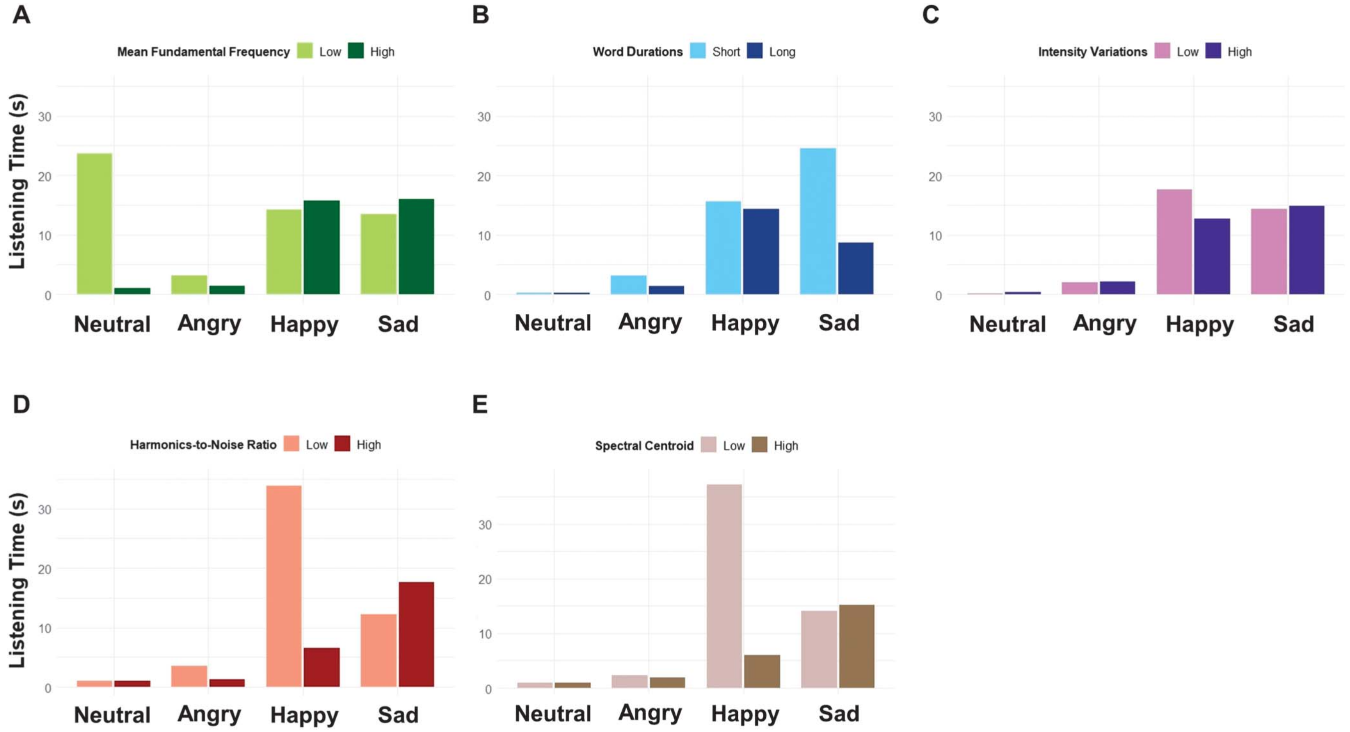


Figure 2. The model predicted listening times to different emotions modulated by (A) mean fundamental frequency, (B) word duration, (C) intensity variation, (D) harmonics-to-noise ratio, and (E) spectral centroid. The 25th and 75th percentiles (from all trial-level acoustic measures) were used as high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



shows how each acoustic variable interacts with infants' listening attention to different emotions.

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 aperiodicity in the speech signal. Infants listened more to the happy prosody with a lower HNR (breathier 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.

Discussion

To better understand early emotional speech perception, we investigated infants' listening attention to happy, angry, sad, and neutral prosodies in spoken words. Notably, we used nonrepeating 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 F0, word duration, intensity variation, HNR, and spectral centroid—to outline their roles in infants' listening attention to emotional prosody.

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 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). Neither less intensity variation nor lower spectral centroid (i.e., less brightness in sounds) is common in a typical happy tone, but they may mimic the soothing voice (Fernald et al., 1984) that infants frequently hear early in life. Although 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 nontypical acoustic constituents in happy prosody.

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,[4] regardless of the speaking styles (IDS or ADS). One major difference was that this study introduced another negative prosody—angry in the stimuli. The current listening task with high affective variations in nonrepeating spoken words per trial 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 et al.'s (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 prosodies. 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) 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 et al.'s (2002) report 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 noise) and spectral centroid (brighter sound) was associated with more of infants' attention 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. Although it may be overstretched to state 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.

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 nonrepeating words, it is surprising that 3- to 12-month-old infants in this 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

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

newborns, the newborns were able to generalize across speakers and form two emotional prosodic categories. Taking the Mastropieri and Turkewitz (1999) together with ours, we believe that infants before the age of 1 year can extract emotional prosodic categories over various nonrepeating 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).

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 ADS we used in this 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 socioemotional 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.

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 this 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 nonrepeating 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. Although 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 socioemotional inputs. Together, we propose that the different emotional processing across male and female individuals may be a product of very large or long-term differences in the learning environments.

Limitations and Future Directions

There are some limitations to this 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 1 year. 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, 5- and 7-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 behavior of emotional speech perception. Third, we used emotional ADS, not the commonly used 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 the infant listener 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.

This study fits into a bigger picture of the interplay between socioemotional 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 socioaffective processing skills and language skills mutually affect one another in this population (Bahn et al., 2021). CIs 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 prosodies (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 socioemotional 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 socioemotional speech input may shape language development in these special populations.

Conclusions

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 the 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 socioaffective development and language learning in typically developing children as well as children with problems in emotional prosody perception.

Acknowledgments

This work was supported by the College of Liberal Arts, University of Minnesota's Graduate Research Partnering Project and the Department of Speech-Language-Hearing Sciences, University of Minnesota's Bryng Bryngelson Fund. We would like to thank Megha Sundara for technical support in setting up the infant laboratory.

References

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

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

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, 64*(3), 993–1007. https://doi.org/10.1044/2020_JSLHR-20-00413

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. https://doi.org/10.1016/j.tics.2007.06.003

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.

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. https://doi.org/10.1016/j.jecp.2010.04.012

Boersma, P., & Weenink, D. (2020). *Praat: Doing phonetics by computer (Version 6.1.09)* [Computer program]. https://www.praat.org

Chong, S. C. F., Werker, J. F., Russell, J. A., & Carroll, J. M. (2003). Three facial expressions mothers direct to their infants. *Infant and Child Development, 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 (Version 1)* [Computer program]. The University of Texas at Austin.

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

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

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

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). Association for Computing Machinery. https://doi.org/10.1145/3243274.3243313

Dupuis, K., & Pichora-Fuller, M. K. (2010). *Toronto Emotional Speech Set (TESS)*. University of Toronto, Psychology Department.

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. https://doi.org/10.1016/S0163-6383(84)80176-9

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), 104–113. 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

Flom, R., & Bahrick, L. E. (2007). The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology, 43*(1), 238–252. https://doi.org/10.1037/0012-1649.43.1.238

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), 429–435. https://doi.org/10.1007/s12576-011-0156-9

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)Talk to me. *Current Directions in Psychological Science, 24*(5), 339–344. https://doi.org/10.1177/0963721415595345

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

Hoemann, K., Xu, F., & Barrett, L. F. (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 D. Gökçay & G. Yildirim (Eds.), *Affective computing and interaction: Psychological, cognitive and neuroscientific perspectives* (pp. 208–243). IGI Global. https://doi.org/10.4018/978-1-61692-892-6.ch010

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

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, 13*(4), 341–348. https://doi.org/10.1002/icd.364

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,* Article 528193. 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

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

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. https://doi.org/10.1111/j.1467-7687.2010.00951.x

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

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

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

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.

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

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

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, 35*(3), 204–214. https://doi.org/10.1002/(SICI)1098-2302(199911)35:3<204::AID-DEV5>3.0.CO;2-V

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

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 *Frontiers in Artificial Intelligence and Applications* (Vol. 265, pp. 3–14). IOS Press. https://doi.org/10.3233/978-1-61499-434-3-3

Mumme, D. L., Fernald, A., & Herrera, C. (1996). Infants' responses to facial and vocal emotional signals in a social referencing paradigm. *Child Development, 67,* 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

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

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

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

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

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

Schröder, M. (2001, September). *Emotional speech synthesis: A review*. Proceedings of the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark.

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

Stern, D. N., Spieker, S., & MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Developmental Psychology, 18*(5), 727–735. https://doi.org/10.1037/0012-1649.18.5.727

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

The 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

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., 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., & 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), 383–403. https://doi.org/10.1037/0033-2909.134.3.383

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

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), 373–377. https://doi.org/10.1037/0012-1649.22.3.373

Walker-Andrews, A. S. (2008). Intermodal emotional processes in infancy. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (pp. 364–375). The Guilford Press.

Walker-Andrews, A. S., & Grolnick, W. (1983). Discrimination of vocal expressions by young infants. *Infant Behavior & Development, 6,* 491–498. 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

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. https://doi.org/10.20982/tqmp.08.1.p052

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,* 136236132199572. https://doi.org/10.1177/1362361321995725

Appendix

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