Maternal Vocal Interactions with Infants: Reciprocal Visual Influences

Sandra E. Trehub1, Judy Plantinga1 and Frank A. Russo2

1University of Toronto
2Ryerson University

Abstract

The present study examined the influence of infant visual cues on maternal vocal and facial expressiveness while speaking or singing and the influence of maternal visual cues on infant attention. Experiment 1 asked whether mothers exhibit more vocal emotion when speaking and singing to infants in or out of view. Adults judged which of each pair of audio excerpts (in view, out of view) sounded more emotional. Face-to-face vocalizations were judged more emotional than vocalizations to infants out of view. Moreover, mothers smiled considerably more while singing than while speaking to infants. Experiment 2 examined the influence of video feedback from infants on maternal speech and singing. Maternal vocalizations in the context of video feedback were judged to be less emotional than those in face-to-face contexts but more emotional than those in out-of-view contexts. Experiment 3 compared six-month-old infants’ attention to maternal speech and singing with audio-only versions or with silent video-only versions. Infants exhibited comparable attention to audio-only versions of speech and singing but greater attention to video-only versions of singing. The present investigation is unique in documenting the contribution of infant visual feedback to maternal vocal emotion in contexts that control for infants’ presence, visibility, and proximity.

Keywords: singing; speech; audiovisual; mothers; infants

Mothers talk and sing to pre-verbal infants to express their feelings and regulate infant attention and arousal (Falk, 2004; Fernald, 1992; Trehub & Trainor, 1998). The heightened positive expressiveness of such infant-directed (ID) vocalizations underlies their apparent efficacy in capturing and maintaining infants’ attention (Corbeil, Trehub, & Peretz, 2013; Fernald, 1993; Singh, Morgan, & Best, 2002; Trainor, Clark, Huntley, & Adams, 1997). For example, infants are more attentive to audio samples of ID than to adult-directed (AD) or non-ID speech and singing.

This research was supported by the Social Sciences and Humanities Research Council of Canada and by AIRS (Advancing Interdisciplinary Research in Singing). Correspondence should be addressed to Sandra E. Trehub, Department of Psychology, University of Toronto Mississauga, Mississauga, ON, Canada L5L 1C6. Email: sandra.trehub@utoronto.ca

© 2015 John Wiley & Sons Ltd
(Fernald, 1985; Masataka, 1999; Trainor, 1996). They also exhibit greater attentiveness and modest increases in positive affect to audiovisual recordings of ID rather than AD speech by female actors (Werker & McLeod, 1989). Moreover, live maternal singing modulates infant arousal, as reflected in changing salivary cortisol levels (Shenfield, Trehub, & Nakata, 2003).

Naïve adult listeners readily distinguish natural ID from non-ID vocalizations (e.g., Bryant & Barrett, 2007; Trainor, 1996; Trehub et al., 1997), which differ in a number of features including pitch level and timing (Fernald, 1989; Fernald & Simon, 1984; Kitamura, Thanavishuth, Burnham, & Luksaneeyanawin, 2002; Trainor et al., 1997). Interestingly, such listeners rarely note pitch or timing as factors underlying their identification of ID versions. Instead, they highlight positive vocal expressiveness such as a warm or loving tone of voice (Trainor; Trehub et al.), which is not readily amenable to measurement.

According to Falk (2004, 2009), the ID style of vocalization arose from ancestral (hominin) mothers’ need to put infants down so they could forage freely. By heightening their vocal expressiveness, mothers were able to keep ‘in touch’ with infants after suspending physical contact. Contemporary caregivers have access to slings or carriers that can free their hands while carrying infants (Hrdy, 1999), but the practice of maintaining almost continuous physical contact with infants is largely restricted to agrarian caregivers in the developing world whose concerns center around safety, survival, and interdependence (Hrdy, 2009; Keller, 2007; LeVine, 1988). In such cultural contexts, vocalizations to infants are typically soothing rather than arousing (Richman, Miller, & Levine, 1992; Trehub & Trainor, 1998), and face-to-face interactions with infants are not accorded particular importance (Kärtnner, Keller, & Yovsi, 2010).

Middle-class mothers in the developed world have very different concerns and socialization goals. They typically provide arousing vocalizations for infants, often without touch, with the goal of promoting social, emotional, and cognitive development (Richman, Miller, & Solomon, 1988; Trehub & Trainor, 1998). These arousing vocalizations commonly occur in face-to-face contexts, with important consequences for infant attention, affect, and learning (Beebe, Alson, Jaffe, Feldstein, & Crown, 1988; Gogate, Bahrick, & Watson, 2000). At times, however, caregivers speak or sing to maintain the composure of infants who are temporarily out of view, as in rear-facing car seats.

In the present investigation, we asked how the emotionality of maternal vocalizations is affected by interacting with infants in or out of view. If maternal vocal emotion is driven primarily by caregiving intentions and infants’ presumed needs (Falk, 2004; Fernald, 1989), then out-of-view infants, by virtue of being more vulnerable to distress, would prompt mothers to exhibit greater vocal emotion. There is considerable evidence, however, of reciprocity and mutual regulation in mother-infant interaction (Cohn & Tronick, 1988; Lavelli & Fogel, 2013). For example, mothers mirror the affective displays of infants (Markova & Legerstee, 2006) who reward mothers with signs of engagement (gaze, facial expression, vocalizations) that promote affectionate maternal vocal and visual responses (Lavelli & Fogel, 2013). Undoubtedly, contingent responsiveness in the face-to-face context contributes to dyadic harmony, but it is unclear whether such reciprocal feedback is necessary for optimizing vocal emotion in speech and song to infants. Because positive vocal tone promotes infant engagement (Corbeil et al., 2013; Singh et al., 2002), it is important to understand how caregivers’ vocal tone is affected by situational factors like the visibility of infants.
In Experiment 1, mothers were recorded speaking and singing to their five-month-old infants in two conditions, one with the infant in full view, another with an opaque curtain between mother and infant. Unlike previous studies in which mothers spoke or sang in an infant’s presence or absence (Fernald & Simon, 1984; Trainor et al., 1997) or in infant’s real or imagined presence (Trehub et al., 1997), mothers in the present study spoke and sang to actual infant listeners who were in view at times and out of view but equally near at other times. As noted, greater emotion-regulatory concerns for out-of-view infants could lead to heightened vocal emotion. Alternatively, continuous feedback from infants in face-to-face interactions could elicit greater vocal emotion from caregivers. Because songs have prescribed words and tunes and mothers perform them for infants in a relatively uniform manner on different occasions (Bergeson & Trehub, 2002), one might expect singing to be unaffected by infants’ visibility so long as they were present. Note, however, that mothers vary their performances of the same songs for infants and preschoolers (Bergeson & Trehub, 1999).

Instead of coding mothers’ speech and singing according to pre-determined criteria, numerous untrained adults judged which version of each pair of sung and spoken excerpts (in-view and out-of-view contexts) sounded more emotional or loving. Comparable judgments of caregivers’ vocal expressiveness and intentions have been used in previous research (e.g., Bryant & Barrett, 2007; Fernald, 1989; Trainor, 1996; Trehub, Unyk, & Trainor, 1993).

In Experiment 2, a different group of mothers and infants participated in similar interactions except for an additional condition in which mothers were able to monitor their infant behind the curtain by means of video feedback. The question of interest was whether such visual monitoring would induce maternal expressiveness comparable to that in the in-view condition.

The availability of recorded maternal speech and singing made it possible to compare their effects on infant attention. Given the choice of audio-only versions of ID speech and singing, infants listen longer to the version that sounds happier, whether speech or singing (Corbeil et al., 2013). Specifically, they listen longer to arousing ID speech than to soothing hummed lullabies and to ID singing than to AD speech, but they listen equivalently to ID singing and speech (Corbeil et al., 2013). When the material is presented audiovisually, they exhibit greater attention to ID singing than to ID speech (Costa-Giomi, 2014; Nakata & Trehub, 2004), raising the possibility that visual cues play an important role.

In Experiment 3, six-month-old infants were presented with audio-only or visual-only excerpts of ID speech and singing from the in-view recordings of Experiment 1, and their responsiveness was assessed with a visual preference procedure that has been used extensively to assess infants’ attention to auditory, visual, and audiovisual stimuli (e.g., Costa-Giomi, 2014; Kim & Johnson, 2013; Trainor, 1996). On the basis of infants’ comparable responsiveness to happy-sounding ID speech and singing (Corbeil et al., 2013) and their greater engagement for ID singing than for ID speech presented audiovisually (Costa-Giomi, 2014; Nakata & Trehub, 2004), we expected them to exhibit greater attention to singing than to speech for visual-only presentation but not for audio-only presentation.

**Experiment 1**

In general, mothers’ speech to infants has higher and more variable pitch, slower rate, greater rhythmicity, and greater dynamic range than does their speech to adults.
(Fernald et al., 1989; Stern, Spieker, & MacKain, 1982), with some variations related to infant age, sex, and culture (Fernald et al., 1989; Grieser & Kuhl, 1988; Kitamura & Burnham, 2003; Papoušek & Hwang, 1991). Maternal singing to infants is higher in pitch, slower in tempo, and more temporally regular than is their usual singing (Nakata & Trehub, 2011; Trainor et al., 1997; Trehub et al., 1997). For naïve listeners, however, the most distinctive aspect of maternal singing is its warm or loving vocal tone (Trainor, 1996; Trehub et al., 1997). Maternal vocal adjustments in speech and song to pre-verbal infants are thought to serve emotion-regulatory and affiliative goals (Fernald, 1991, 1992; Kitamura & Burnham; Trehub & Trainor, 1998), a view consistent with their use in infant-present but not infant-absent contexts and in some affectively charged contexts with adults (Trainor, Austin, & Desjardins, 2000). What is unclear, however, is the role of infant feedback in eliciting or sustaining mothers’ positive emotional tone.

There are demonstrable effects of infant visual feedback on maternal vocal quality. For example, manipulations of infant video feedback to reinforce mothers’ high-pitched speech increase the incidence of such speech (Smith & Trainor, 2008). Moreover, mothers’ vocal behavior differs when video feedback from infants is delayed rather than being presented in real time (Braarud & Stormark, 2008; Murray & Trevarthen, 1986).

In the present experiment, mothers interacted with infants by means of speech and singing in two contexts, one in which mother and infant faced one another, the second with both parties in the same position but separated by an opaque screen. In previous studies, naïve adult listeners were sensitive to voice quality differences in infant-present and infant-absent contexts (e.g., Bryant & Barrett, 2007; Trainor, 1996; Trehub et al., 1993, 1997). Here, we asked whether naïve adult listeners would detect differences in maternal vocal tone while speaking and singing with or without infant visual feedback. We considered naïve listeners preferable to trained listeners for the present purposes. For one thing, there are no empirically derived criteria for qualities as amorphous as emotional or loving tone of voice and, therefore, no informed basis for training listeners. Accordingly, intuitive but statistically reliable judgments from many untrained listeners were considered more appropriate than judgments from two or three listeners who relied on specific acoustic features.

**Method**

**Participants.** The participants were 21 culturally and ethnically diverse college students (17 women, 4 men) from an introductory psychology course who were unselected with respect to music background, experience with infants, or gender. (There were many more women than men in the course.) They received partial course credit for their participation. Procedures for recruiting and testing participants were approved by the Research Ethics Board of the University of Toronto, as were the procedures for other experiments.

**Apparatus and Stimuli.** Stimulus recordings consisted of talking and singing interactions of 21 mothers and their five-month-old infants (9 girls, 11 boys, 22.6–26 weeks of age; $M = 24.3$ weeks, $SD = 1.03$). Mothers were from middle-class backgrounds in a culturally and ethnically diverse community surrounding the campus. Each talking and singing interaction was 1 min in duration and occurred under two
conditions, one with the infant in full view (in-view condition) and the other with mother and infant in the same location but with the infant obscured from view by an opaque curtain (out-of-view condition). Recordings from a further five participants were excluded because of infant fussing (n = 2), mothers’ failure to sing (n = 1), or experimenter error (n = 2).

Recordings were made in a double-walled sound-attenuating booth (Industrial Acoustics Corporation 110766, 3 m × 2.5 m). Two video cameras (Sony Handycam HDR-XR500V) in the booth were connected to an Edirol LVS-400 Videomix/liveswitcher outside the booth that was connected to an iMac computer, enabling the experimenter to monitor both cameras simultaneously. One camera was focused on the mother and the other on the infant. An opaque curtain hung from the ceiling of the sound booth between mother and infant for the out-of-view condition. The curtain had a colorful print on the side facing the infant but was solid purple on the side facing the mother. High quality audio recordings were obtained by having mothers wear a Sony Electret condenser lapel microphone (ECM-55B) linked to a SoundBlaster X-Fidelity sound card in a Dell (Pentium 4, CPU) computer outside the booth. ARWizard audio recording software was used to create the audio files (bit rate of 44.1 KHz, .wav format). During the recording phase, the infant was seated securely in a high chair in the booth facing the seated mother. The mother was instructed to refrain from touching her infant during the recording session. She was told to sing and talk to her infant as she usually did at home despite the constraints of the situation (i.e., no touching, infant in or out of view). Mothers chose to engage in lively singing and their speech was predominantly lively as well. The experimenter monitored the session from a computer outside the booth, entering the booth between conditions to place or remove the curtain or to tell mothers to switch from singing to talking or vice versa. The first interaction modality (talking, singing) and condition (in view, out of view) were counterbalanced across mother-infant pairs.

Excerpts of the first 6–12 s of each audio recording that were free of infant sounds were extracted with QuickTime Pro and edited to yield a full utterance or musical phrase. There were 42 stimulus pairs, consisting of 21 excerpts of singing in view and out of view from the same mother and 21 pairs of talking in and out of view from those mothers. Duration of sound clips varied slightly between mothers, but was roughly comparable within pairs from each mother. Stimulus pairs were amplitude normalized by means of Sonic Foundry Sound Forge 6.0.

Testing occurred in a double-walled sound-attenuating booth (Industrial Acoustics Corporation) 2.90 m by 2.44 m in size. Stimulus presentation and response tabulation were controlled by a custom-designed program (RealBasic) on a Dell Pentium 4 Windows XP workstation with SoundBlaster X-Fi Fatality sound card. Sounds were delivered through a Harmon/Kardon 3380 amplifier located outside the booth that was connected to two speakers (Audiological GSI) in the sound booth located on either side, approximately 1.5 m from the participant. Participants controlled the presentation of stimuli and entered their responses by means of a Targus Number keypad linked to the computer outside the booth.

Procedure. All participants heard the speech and singing excerpts of all 21 mothers in both contexts (in and out of view). On each trial, they heard a pair of excerpts from each mother (with in- or out-of-view infant) separated by 2-s silence. They received the following written instructions: ‘You will hear pairs of excerpts in
which mothers are either singing to their infant or talking to their infant. After hearing each pair, you will be asked to judge which of the two excerpts sounds more emotional or loving. When the instructions were reiterated verbally, participants were told to indicate which excerpt (first or second) sounded more emotional or loving by selecting 1 (first) or 2 (second) on the keypad. Participants had the option of listening to either clip or both clips again before responding. The experimenter initiated the first trial from outside the booth, with the participant initiating the remaining trials. Stimulus pairs were presented in random order, with the first interaction modality (singing, talking) and first condition (in view, out of view) counterbalanced across pairs.

Results

Listeners judged the in-view excerpts as more emotional significantly more often than would be expected by chance for both singing (M singing in view = 14.8, SD = 3.19 vs. M singing out of view = 6.19, SD = 3.19) $t(20) = -6.20$, $p < .001$, and talking (M talking in view = 13.43, SD = 2.89 vs. M talking out of view, $M = 7.57$, $SD = 2.89$), $t(20) = -4.642$, $p < .001$ (see Figure 1).

Video-recordings of mothers made it possible to examine maternal facial expressions, smiling in particular, that might reflect maternal feelings and intentions. The videos, recorded in iMovieHD, were converted to QuickTime movies for the present purposes. Video recordings from 20 mothers (one mother excluded because of obscured face) were edited to provide separate videos of in-view and out-of-view speech and singing from each mother. An independent rater tabulated the duration of maternal smiling in all conditions from the 60-s videos played silently. The ratings were highly correlated (.8) with those of a second rater who independently coded a substantial subset of the videos. Discrepancies between raters arose from minor differences at the boundary between smiling and non-smiling, for example, a very slight smile vs. a neutral face. Durations of smiling across interaction types

Figure 1. Mean ‘More Loving’ Responses for In-view and Out-of-view Versions of Maternal Talking and Singing in Experiment 1.

Note: Error bars are standard errors.

© 2015 John Wiley & Sons Ltd Social Development 25, 3, 2016
(talking, singing) and conditions (in view, out of view) are shown in Figure 2. Not surprisingly, mothers smiled more when infants were visible. They smiled comparably during speech and singing episodes to infants who were out of view, but they smiled significantly more when singing than when talking to their in-view infant, $t(19) = 4.46, p < .001$ (see Figure 2).

The findings indicate that having infants within close range but out of sight is insufficient to prompt the full complement of maternal adjustments in ID speech and singing. Although mothers might be expected to compensate for their inability to transmit expressive visual cues by enhancing their vocal expressiveness, that was not the case. Instead, adult listeners rated maternal vocal behavior as less emotional or loving when infants were out of view, raising the possibility that visual cues from infants elicit or sustain enhanced vocal emotion. This question was examined further in Experiment 2.

**Experiment 2**

As noted, lesser vocal emotion when infants are out of view may stem from the absence of infant visual feedback. If such feedback is entirely responsible for the discrepancy in maternal emotional expressiveness between the in-view and out-of-view conditions, then the provision of video feedback could reduce or eliminate the differences. In the present experiment, a different group of mothers sang and spoke to infants under the two conditions of Experiment 1 (in view and out of view) and under a third condition in which video feedback was provided from infants behind the opaque curtain. If aspects of the face-to-face context, such as contingent bidirectional feedback (Braarud & Stormark, 2008; Murray & Trevarthen, 1986), have favorable effects on mothers and optimize coordinated interaction, then mothers might speak and sing with greater emotionality in the in-view than in the video-view condition.
Method

Participants. The participant raters were 18 college students (13 women, 6 men) from an introductory psychology course who were recruited as in Experiment 1. They received partial course credit for their participation.

Apparatus and Stimuli. Stimulus recordings consisted of talking and singing interactions from 22 mothers and their five-month-old infants (8 girls, 15 boys, 21.7–25.4 weeks of age; \( M = 23.1 \) weeks, \( SD = 1.25 \)) from the same community as the participants in Experiment 1. Audio recordings from an additional eight mothers were excluded because of experimenter error (\( n = 2 \)) or infant crying (\( n = 6 \)). The apparatus was the same as that in Experiment 1 except that a third camera (Sony Digital Handycam DCR-TRV510 NTSC) focused on the infant could transmit a video display of the unseen infant by means of a portable computer (MacBook Pro 2.1).

The recording procedure was the same as in Experiment 1 except for the additional condition during which mothers sang or talked to their infant behind the curtain while watching the infant in real time on the portable computer. The three conditions of the experiment were as follows: (1) infant in view, (2) infant behind opaque curtain (out of view), and (3) infant behind the curtain but visible on the digital display (video view). First condition (out of view and video view) and first interaction modality (singing, talking) were counterbalanced across mother-infant dyads. Because the comparison of principal interest involved mothers communicating with out-of-view infants vs. infants visible on video, the in-view condition was always recorded last.

QuickTime Pro was used to create audio excerpts of 6–12 s from the initial portions of the recordings for each condition and mother. There were 80 pairs of excerpts, 20 of each type: singing in view vs. out of view, singing in view vs. video view, talking in view vs. out of view, and talking in view vs. video view (4 each from 18 mothers, and talking-only or singing-only pairs from another 4 mothers). Durations of audio excerpts varied slightly across mothers but were comparable for each mother. For the few maternal sound files that contained occasional positive infant vocalizations, those vocalizations were copied and added to the comparison excerpts (same mother-infant dyad) to preclude their influence on adults’ judgments. No excerpts had negative infant vocalizations. In other words, paired comparisons for each condition had either no infant vocalizations or the same occasional infant vocalizations in each excerpt. Excerpts in each pair were amplitude normalized with Sonic Foundry Sound Forge 6.0.

Testing occurred in the sound-attenuating booth, with stimulus presentation and response tabulation controlled by custom software on the same Dell workstation as Experiment 1. As in Experiment 1, participants controlled stimulus presentation and entered their responses on a Targus Number keypad.

Procedure. The procedure was the same as in Experiment 1. Pairs were presented randomly with first interaction modality (singing, talking) and first condition (in view, out of view, video view) counterbalanced across pairs.

Results and Discussion

Separate repeated measures analyses of variance (ANOVAs) for the singing and talking conditions, with context (infant in view, out of view, or on video) as a within-subjects factor and times selected as more emotional as the dependent
measure revealed a significant effect of context for singing $F(3, 51) = 55.456$, $p < .001$, $\eta^2_p = .765$, and for talking, $F(3, 51) = 107.955$, $p < .001$, $\eta^2_p = .864$. Planned comparisons (Bonferroni corrected) revealed that listeners more frequently judged in-view excerpts as more emotional than out-of-view excerpts for singing ($M = 13.94$, $SD = 13.94$ and $M = 6.06$, $SD = 2.10$, respectively), $t(17) = 7.970$, $p < .001$, and for talking ($M = 16.17$, $SD = 2.07$ and $M = 3.83$, $SD = 2.07$, respectively), $t(17) = 12.669$, $p < .001$ (see Figure 3, left panel), replicating the findings of Experiment 1. In addition, adults more frequently judged in-view excerpts as more emotional than video-view excerpts for singing ($M = 13.39$, $SD = 2.09$ and $M = 6.61$, $SD = 2.09$, respectively), $t(17) = 8.857$, $p < .001$, and talking ($M = 14.14$, $SD = 2.23$ and $M = 5.83$, $SD = 2.23$, respectively), $t(17) = 7.929$, $p < .001$ (see Figure 3, right panel). The findings revealed that continuous video feedback from infants did not generate maternal performances that matched the expressiveness of those in the in-view condition. What remained unclear, however, was whether video feedback made any contribution to maternal expressiveness. This question was examined in a follow-up experiment.

**Experiment 2a**

Ratings of greater maternal expressiveness for the in-view than video-view conditions in Experiment 2 ruled out the visibility of infants as the factor underlying differential expressiveness in the in-view and out-of-view contexts. It is possible, however, that video feedback from infants could enhance maternal expressiveness relative to the absence of such feedback. To that end, adult listeners judged which of the paired excerpts of maternal speech and singing from video-view and out-of-view conditions sounded more emotional.

**Method**

**Participants.** The participants were 18 adults (15 women, 3 men) who received token payment for their participation.

**Apparatus and Stimuli.** The apparatus was the same as in Experiment 2. The stimuli from Experiment 2 were used to create two different pairs for each mother, one involving out-of-view excerpts of singing and talking, the other involving video-view excerpt of singing and talking.

**Procedure.** The procedure was the same as in Experiment 2.

**Results and Discussion**

Separate repeated measures ANOVAs on the singing and talking conditions evaluated the effect of context (video view, out of view) on listeners’ choice of the more emotional excerpt in each pair. These analyses confirmed that the differences between contexts were significant for singing, $F(1,17) = 7.766$, $p = .013$, $\eta^2_p = .314$, and for speech, $F(1, 17) = 4.620$, $p = .046$, $\eta^2_p = .214$ (see Figure 4). Adults more frequently rated the video-view excerpts as more emotional than the out-of-view excerpts for singing ($M =11.61$, $SD = 2.453$ vs. $M = 8.39$, $SD= 2.453$) and for speaking ($M = 11.11$, $SD = 2.193$ vs. $M = 8.89$, $SD = 2.193$).
Figure 3. Mean ‘More Loving’ Responses for In-view and Out-of-view Versions (Left Panel) and In-view and Video-view Versions (Right Panel) of Maternal Talking and Singing in Experiment 2.

Note: Error bars are standard errors.
Video recordings were edited into four segments: talking and singing with infant out of view, and talking and singing with video view of infant. Video recordings from 25 mothers, which provided a clear facial view of the mother for the entire 60-s session, were coded for cumulative seconds of smiling in out-of-view and video-view contexts for singing and speech. Mothers smiled significantly more in the video-view context, when they could see infants on the monitor (\(M = 15.98, SD = 18.68\)), than in the out-of-view context when infants remained unseen (\(M = 7.38, SD = 13.21\), \(t(49) = 3.001, p = .004\), even though smiling in the former context had no communicative significance. Instead, the sight of the infant seemed to intensify maternal feelings, resulting in greater smiling and greater vocal emotion. Greater maternal smiling for singing than for speech, although evident in the in-view condition of Experiment 1, was not evident under the artificial circumstances of infants behind a curtain or with video feedback. The implications of ID auditory and visual displays for infant listeners and viewers were examined in Experiment 3.

**Experiment 3**

Although infants’ relative responsiveness to ID and non-ID speech (e.g., Cooper & Aslin, 1990; Fernald, 1985) and their responsiveness to ID and non-ID singing have been well-documented (Masataka, 1999; Trainor, 1996), there have been few comparisons of their responsiveness to ID speech and singing. In the two published studies with audio-only materials, infants exhibited comparable attention to sung and spoken portrayals of children’s songs (Corbeil et al., 2013; Costa-Giomi, 2014). The two published studies with audiovisual materials, one involving 6-month-olds and natural maternal speech and singing (Nakata & Trehub, 2004), the other involving 11-month-olds and portrayals of ID speech and singing (Costa-Giomi, 2014), found greater infant attention to singing than to speech. Surprisingly, 11-month-olds did not respond preferentially to visual-only versions of ID singing (Costa-Giomi, 2014), which does not preclude the possibility that infants would do so for natural

---

**Figure 4.** Mean ‘More Loving’ Responses for Out-of-view and Video-view Versions of Maternal Talking and Singing in Experiment 2a.

*Note:* Error bars are standard errors.
versions rather than portrayals in the absence of an infant. Mothers produce distinctive facial expressions (Chong, Werker, Russell, & Carroll, 2003) and head movements (Smith & Strader, 2014) when they speak to infants, and Experiment 1 revealed that they smile more when singing than when talking to infants. It is possible, then, that infants’ greater interest in audiovisual renditions of maternal singing (Nakata & Trehub, 2004) stems, in part, from more positive facial displays during such singing.

In the present experiment, we compared infants’ responsiveness to audio-only samples of maternal speech and singing and to visual-only samples of maternal speech and singing, all recorded in face-to-face contexts. On the basis of infants’ comparable responsiveness to happy-sounding (audio-only) ID speech and singing (Corbeil et al., 2013; Costa-Giomi, 2014), their greater responsiveness to audiovisual samples of ID singing than speech (Costa-Giomi, 2014; Nakata & Trehub, 2004), and greater maternal smiling when singing than when speaking to infants (Experiment 1), we predicted that infants would show greater attention to visual-only excerpts of singing but comparable attention to audio-only samples of speech and singing from Experiment 1. In other words, we expected to replicate the previous ID speech and singing comparisons based on auditory cues alone (i.e., no difference) but to find an advantage for ID singing based on visual cues alone. Just as happy-sounding vocalizations have favorable consequences for infant attention (Corbeil et al., 2013; Singh et al., 2002), happy-looking faces may have comparable consequences.

Method

Participants. The participants were 64 full-term infants (31 boys, 33 girls) 5.5–6.5 months of age (23.4–30.6 weeks, \( M = 26.6 \)) from the same community as the infant participants in Experiments 1 and 2. According to parents’ report, infants had no history of ear infections or family history of hearing loss. An additional 16 infants were excluded because of experimenter error (\( n = 5 \)), infant crying (\( n = 3 \)), infant moving out of camera range (\( n = 2 \)), infant failure to look at the target visual stimulus (\( n = 3 \)), and parents’ interaction with infants during the test session (\( n = 3 \)).

Apparatus and Stimuli. Testing took place in a double-walled sound-attenuating booth (Industrial Acoustics Corporation 110766, 3 m \( \times 2.5 \) m) with dim lighting. There were three monitors: one directly in front of the infant at a distance of 1 m, one 45° to the infant’s right, and the other 45° to the left. The presentation of stimuli and the recording and tabulation of looking times were controlled by a custom program (RealBasic) on a Windows 7 workstation. Sounds were delivered through an amplifier (Harmon/Kardon 3380) located outside the booth connected to two loudspeakers (Audiological GSI) located directly below each of the side monitors inside the booth. A Sony Handycam HDR-XR500V was connected to an Edirol LVS-400 Videomix/liveswitcher outside the booth that was connected, in turn, to an iMac 10.1 display.

Half of the infants were assigned to an audio-only condition and half to a visual-only condition. The audio stimuli consisted of two excerpts (one singing, one talking) from each of four mothers in the in-view condition of Experiment 1. The excerpts of maternal speech and singing contained no infant sounds and ranged from 19 to 31 s in duration across mothers (but comparable duration within mothers). The visual stimuli were silent video excerpts from the same mothers, with a full view of the mother’s face in each case.
Procedure. Infants were tested with the headturn preference procedure in which looking at one locus or another is linked to one of two contrasting stimuli. Such looking becomes an index of listening when the contrasting stimuli presented at those loci are auditory. The infant sat on the parent’s lap facing the central monitor. Parents listened to music over headphones to prevent them from hearing any stimuli presented to infants. They were instructed to refrain from interacting with infants except to return them to the original position, if necessary. An experimenter seated outside the booth observed the infant on the iMac computer (without sound), which displayed images from the camera in the booth. She depressed one key to indicate when the infant was looking at the flashing screen and another to indicate when the infant looked away from that monitor. Custom software automatically tabulated infant looking time for each trial and cumulative looking time for each stimulus type (audio or silent video samples of ID singing or speech). The experimenter was unaware of the stimuli being presented on any trial. Each trial began with the central screen flashing red. When the experimenter indicated that the infant was fixated on the central monitor, that screen stopped flashing and a side screen began flashing. In the audio-only condition, when the infant fixated on the side monitor, a high-resolution image of a checkerboard with central red dot appeared at that locus, and an auditory stimulus (e.g., ID talking) was presented. The stimulus repeated until the infant looked away from the monitor for 2 s, at which time the sound stopped and the visual display disappeared, leaving a black screen. The center screen flashed to refocus infants’ attention at midline. Once central fixation was achieved, flashing began on the screen on the opposite side. Fixation on that screen initiated the same checkerboard pattern in conjunction with the contrasting auditory stimulus (e.g., ID singing). Again, the visual and auditory stimuli continued until the infant looked away for 2 s. Infants in the visual-only condition were presented with a silent video of a mother singing on one side and a silent video of the same mother talking on the other side. Talking and singing stimuli were presented on alternating trials, as for the audio-only procedure. For the auditory-only and visually-only conditions, presentation of the target stimuli was contingent on infants’ visual fixation of the relevant monitor, so that looking times at each monitor provided an index of infants’ relative attention to each of the comparison stimuli. There were 12 trials in the audio-only and visual-only conditions, 6 talking and 6 singing. The location of singing and talking stimuli was consistent for individual infants, but first presentation (left or right) and first stimulus (singing or talking) were counterbalanced across infants.

Results and Discussion

Proportion of looking time during audio-only versions of speech ($M = .460, SD = .24$) and singing ($M = .540, SD = .24$) did not differ, $t(31) = .929$. By contrast, proportion of looking to visual-only versions of singing ($M = .575, SD = .14$) was significantly greater than that for speech ($M = .425, SD = .14$), $t(31) = 3.078$, $p = .004$, two-tailed (see Figure 5). These findings highlight the contributions of emotive visual gestures to infants’ engagement with singing, which may account for greater infant attention to audiovisual renditions of maternal singing than to comparable renditions of maternal speech (Nakata & Trehub, 2004).
General Discussion

The present study explored the effects of situational factors, in particular, the presence or absence of infant visual feedback, on maternal speech and singing to five-month-old infants. Mothers’ speech and singing was characterized by greater emotionality during face-to-face interactions than during interactions in which infants were obscured from view (Experiment 1). This finding is consistent with more emotive speech and singing in infants’ presence than in their absence (Fernald, 1989; Trainor, 1996; Trehub et al., 1993, 1997). Note, however, that mothers and infants in the current study were in the identical location whether infants were in or out of view.

Our finding of greater vocal emotion in face-to-face contexts is inconsistent with such behavior being driven primarily by mothers’ caregiving intentions (Falk, 2004; Fernald, 1989; Kitamura & Burnham, 2003; Stern et al., 1982). Presumably, mothers in the present study sought to maintain infants’ composure in both contexts. In fact, it is likely that they had greater concerns when their infants were out of view, but that context resulted in reduced rather than enhanced vocal emotion. Nevertheless, Falk’s notion of maternal vocalization compensating for the absence of tactile contact is consistent with more soothing maternal vocalization in cultures where proximal interaction prevails and more arousing vocalization in cultures where distal interaction prevails (Trehub & Trainor, 1998).

To further investigate the effect of infant feedback on maternal behavior, Experiment 2 had an additional condition in which mothers received one-way video feedback from infants. Mothers’ vocalizations were more emotive when infants were available face to face than when they were visible on a monitor. Inspection of the videos of infants during maternal interactions revealed differences in infant behavior in face-to-face and video contexts. When mothers and infants faced one another, infants’ engagement was clearly evident from their extended visual fixations on the mother’s face, often accompanied by eye widening. It is likely that infants in the video-view condition were engaged by mothers’ singing, but they often gazed at the patterns on the curtain rather than in the direction of the camera (and mother), providing ambiguous visual feedback for mothers.

Figure 5. Mean Proportion of Infant Looking Time for Audio-only and Visual-only Samples of Maternal Talking and Singing.
Note: Error bars are standard errors.
In addition, the video context provided a partial view of infants (head and shoulders), in contrast to the full view available in the face-to-face context, which may have provided further cues to infants’ engagement. In principle, reciprocal video feedback for mothers and infants, as in Braarud and Stormark (2008), could result in maternal vocal behavior that approximates the emotionality observed in face-to-face contexts. However, the difficulty of establishing bidirectional eye contact under such circumstances makes that outcome unlikely.

Despite the limitations of video feedback, maternal vocalizations were more emotive in that situation than when infants were entirely out of view. If still images of an infant activate reward-processing areas in the mother’s brain (Strathearn, Li, Fonagy, & Montague, 2008), perhaps it is not surprising that dynamic visual images in the present study led to more maternal smiling and greater vocal emotion during video-view than out-of-view episodes. As expected, mothers made greater use of smiling, an obvious display signal, when they spoke or sang in face-to-face contexts than in other contexts. What is especially notable is that they smiled almost continuously when they sang to infants in view and only intermittently when they talked in that context. It is unlikely that greater maternal smiling during singing episodes arose from greater infant smiling. Infants’ characteristic response to maternal singing is intense engagement, as reflected in prolonged visual fixation of the mother’s face and reduced body movement (Nakata & Trehub, 2004). In any case, it is likely that mothers’ smiling affected their vocal tone. Smiling alters the shape of the vocal tract, leading to speech that listeners characterize as happy-sounding (Tartter, 1980; Tartter & Braun, 1994). As a result, smiling to in-view infants not only reflected mothers’ feelings but also helped convey those feelings vocally as well as visually.

There is increasing consideration of the facial gestures that accompany maternal speech to pre-verbal infants (Chong et al., 2003; Kim & Johnson, 2013, 2014; Smith & Strader, 2014). In addition to distinctive facial expressions that include exaggerated smiles and mock surprise (Chong et al., 2003), mothers exhibit more head movement during ID than AD speech, with such movement linked to speech prosody (Smith & Strader, 2014). To date, however, there has been no consideration of the facial expressions of maternal singers even though the face-to-face context is more common for maternal singing than for maternal speech, the latter often involving directed attention to objects and events in the environment. It is notable, then, that mothers in the present study smiled more while engaged in face-to-face singing rather than talking. The practice of singing the same songs repeatedly to infants and performing them in a highly stereotyped manner (Bergeson & Trehub, 2002; Trehub et al., 1997) may make singing less cognitively resource intensive than speaking. Fewer cognitive demands could facilitate greater visual expressiveness during singing than during speech, contributing to the efficacy of maternal singing for infant emotion regulation (Ghazban, 2013; Trehub, Ghazban, & Corbeil, 2015).

Studies of ID speech and singing have focused primarily on acoustic cues or consequences for infant listeners (e.g., Fernald, 1985; Trainor, 1996; Trainor et al., 1997; Trehub et al., 1997). It has become clear, however, that infants prefer silent ID to AD faces of talkers, but they prefer happy to sad faces of talkers regardless of the ID or AD register (Kim & Johnson, 2013), just as they prefer happy speech and singing to neutral or soothing versions (Corbeil et al., 2013; Singh et al., 2002). In addition, three- and five-month-old infants look longer at silent displays of ID talking than at simultaneous (side-by-side) silent displays of AD talking whether they watch the displays in silence, while listening to ID speech, or while listening
to AD speech (Kim & Johnson, 2014). Although audio samples of ID speech recruit infants’ attention more effectively than AD speech (Cooper & Aslin, 1990; Fernald), ID faces dominate their attention, even in the presence of voices.

Greater infant attention to maternal singing than to speech presented audiovisually (Nakata & Trehub, 2004) may stem from emotive visual displays rather than distinctive melodic and rhythmic features. In the present study, six-month-olds responded no differently to audio samples of maternal speech and singing, in line with previous findings involving ID speech and singing portrayals (Corbeil et al., 2013; Costa-Giomi, 2014). Note, however, that they exhibited greater attention to visual-only versions of maternal singing than speech, perhaps because the former featured more smiling. Face-to-face singing also involves some rhythmic movement, providing a more dynamic visual display than that afforded by face-to-face talk. In previous research, 11-month-olds showed comparable attention to visual-only versions of ID speech and singing that involved portrayals rather than naturally occurring communication (Costa-Giomi, 2014).

Live, face-to-face maternal singing modulates the arousal levels of non-distressed infants (Shenfield et al., 2003). For distressed infants, live maternal singing has greater efficacy than maternal speech for reducing arousal and negative vocalizations (Ghazban, 2013). Because of infants’ ability to detect cross-modal correspondences in ID speech and singing (Trehub, Plantinga, Brcic, & Nowicki, 2013), simultaneous visual cues may highlight the acoustic differences between ID speech and singing, contributing to the arousal-regulatory efficacy of ID singing.

In short, face-to-face contexts enhanced the emotionality of maternal speech and singing, as judged by naïve adult listeners. Maternal vocal emotion is influenced by mothers’ emotion regulatory intentions (Falk, 2004, 2009) and their temporal coordination with infants (Braarud & Stormark, 2008; Beebe et al., 1988), which depends on the reciprocal feedback available in face-to-face contexts. The distinctive vocal tone in such situations raises questions about the widespread use of scripted ID portrayals to evaluate infants’ responsiveness to speech and singing (e.g., Cooper & Aslin, 1990; Corbeil et al., 2013; Costa-Giomi, 2014; Singh et al., 2002). More emotive visual displays when mothers sing than when they speak to infants and their consequences for infant attention also raise questions about the characteristic focus on auditory aspects of ID speech and singing.

Although participating families in the present study were relatively advantaged (i.e., middle-class), the findings have implications for less advantaged families. Infants and toddlers from families with low socioeconomic status (SES) have less speech directed to them than those from higher SES backgrounds (Hart & Risley, 1995; Hoff, 2006), with such differences in input predicting vocabulary and language processing by 18 months of age (Fernald, Marchman, & Weisleder, 2013). Those studies focus on the auditory components of maternal interaction, but it is important to consider the complementary visual components that scaffold the acquisition of language (Gogate, Maganti, & Bahrick, 2015) as well as social and emotional development (Kopp, 1989; Lavelli & Fogel, 2013).

References


