Infants Detect Cross-modal Cues to Identity in Speech and Singing

Sandra E. Trehub, Judy Plantinga, and Jelena Brcic

Department of Psychology, University of Toronto, Mississauga, Ontario, Canada

Little is known about infants' perception of cross-modal cues to identity, but the importance of recognizing familiar individuals makes it likely that this skill would be evident early in life. Infants 6–8 months of age were tested on their ability to link dynamic cross-modal cues to the identity of unfamiliar speakers and singers. After exposure to speech or singing, infants watched two silent videos, one featuring the previously heard speaker or singer. Infants looked significantly longer at the video of the person heard previously, which indicates that they can match auditory and visual cues to the identity of unfamiliar persons.

Key words: **infants; speech; singing; perception; cross-modal; identity**

Introduction

Whereas traditional accounts of voice recognition consider person-specific (i.e., indexical) aspects of speech (e.g., unique voice quality) as entirely separate from segmental (phonetic) and prosodic aspects,[1] more recent views invoke interdependent processing of these cues both within and across modalities.[2] After modest exposure and practice, adults recognize familiar voices from sine wave analogues of speech that eliminate timbre cues while preserving cues to dynamic articulatory gestures.[3] In addition to the dynamic vocal gestures that provide a unique acoustic signature, spatial frequency cues from head and lip movements provide a unique visual signature.[4]

During face-to-face conversations, listeners have access to visual as well as vocal aspects of the speaker's communication. Visual features increase the efficacy of communication, especially in adverse listening conditions. Movements of the lips are of paramount importance, but highly controlled conditions of the laboratory have revealed other less obvious visual cues that make significant contributions to speech processing. For example, natural head motion enhances the intelligibility of speech.[5]

In recent years interest has been increasing concerning possible correlations between the dynamic vocal and visual signatures that have been identified. In fact, adults are capable of matching auditory and visual cues to the identity of unfamiliar individuals when the speech samples are presented in normal, but not in reversed order,[6] which implies that common temporal cues are critical. Specifically, adults exposed to single utterances followed by silent visual displays of two unfamiliar speakers matched the correct video to the previously heard speaker at modest but above chance levels (60% correct). Adults performed similarly when they received a silent visual display followed by utterances from two different speakers. Although performance is optimized by comprehension of the language in question, adults sometimes perceive relations between dynamic auditory and visual cues from a foreign language.[2]

Although children are sensitive to cues to vocal identity, they require more cues than adults do for successful identification. Preschoolers recognize the voices of classmates[7] and of familiar cartoon characters.[8] Within a few days of

Address for correspondence: Sandra E. Trehub, Department of Psychology, University of Toronto Mississauga, Mississauga, Ontario, Canada L5L 1C6. sandra.trehub@utoronto.ca

The Neurosciences and Music III: Disorders and Plasticity: Ann. N.Y. Acad. Sci. 1169: 508–511 (2009).
doi: 10.1111/j.1749-6632.2009.04851.x

birth, infants differentiate their mother's voice from a stranger's voice.[9] They also differentiate dynamic images of their mother from those of a stranger.[10] Although little is known about infants' perception of cross-modal cues to identity, the importance of person recognition for survival (e.g., distinguishing friend from foe) makes it likely that infants would have mechanisms for the efficient extraction of identity cues across modalities. To this end, Trehub and Brcic[11] and Trehub, Plantinga, and Brcic (manuscript in preparation) explored infants' ability to perceive relations between dynamic auditory and visual cues from unfamiliar speakers and singers.

Methods

These investigators used auditory and visual stimuli from mother–infant interactions because of infants' well-documented preference for infant-directed over noninfant-directed speech and singing and their greater emotional transparency.[12] The presumption was that cues to identity would also be more transparent in infant-directed speech and singing. The speech materials consisted of 30-s audio recordings from natural interactions of six different women with their 6-month-old infants and 30-s silent video-recordings of the same dyads. The auditory and visual recordings from the same women involved different (i.e., non-matching) speech samples.

Each of 48 infants heard a 30-s sample of infant-directed speech from one woman, after which they were tested with two silent videos, including one from the previously heard speaker. The speakers were paired on the basis of similar visual features (e.g., hair color and length). Half of the infants tested with any pair of silent videos were familiarized with the voice of one woman of the pair; half were familiarized with the other woman's voice. On a series of test trials, each video was presented until the infant looked away. Cumulative looking time provided an index of infants' interest in the person depicted in each video. If infants perceived the two women as equally unfamiliar, they should look equally long at both images. If infants perceived one woman as being more familiar on the basis of the vocal cues heard previously, they should exhibit differential looking at the dynamic images.

For the corresponding singing comparisons, 6- to 8-month-old infants were exposed to an audio-recording of a mother singing a song to her infant, after which they watched silent videos of two unfamiliar women—the previously heard singer and another person—singing infant-directed versions of another song. In a subsequent experiment, the auditory and visual materials were derived from different portions (e.g., first half, last half) of the same song.

Results

Infants looked significantly longer at the video of the previously heard speaker (57.3% of total looking) than would be expected by chance (50%), $t(47) = 3.64$, $P < 0.001$ (Fig. 1), which implies that the infants matched dynamic auditory and visual cues to speaker identity. Adults exposed to the same familiarization materials followed by 7-s samples from the paired video-recordings performed well above chance levels (Fig. 2). When the auditory and visual materials involved different songs, infants looked equally long at both videos, revealing no evidence that they perceived auditory and visual cues to the singer's identity (Fig. 1). Adults tested with the same task also failed to discern cross-modal cues to the singer's identity (Fig. 2). Unlike trained singers, who develop a unique style that is recognizable to their fans, untrained singers may vary their performing style across songs, obscuring potential cues to identity. When auditory and visual materials were drawn from different portions of the same song, infants looked significantly longer at the previously heard singer (mean = 57.8%) than

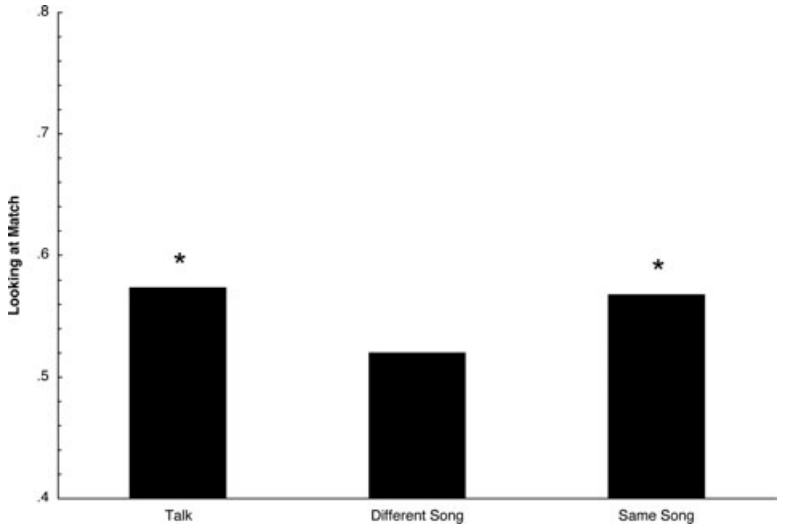

Figure 1. Proportion of time looking at silent video of previously heard speaker or singer. *Indicates responses significantly exceeding chance levels.

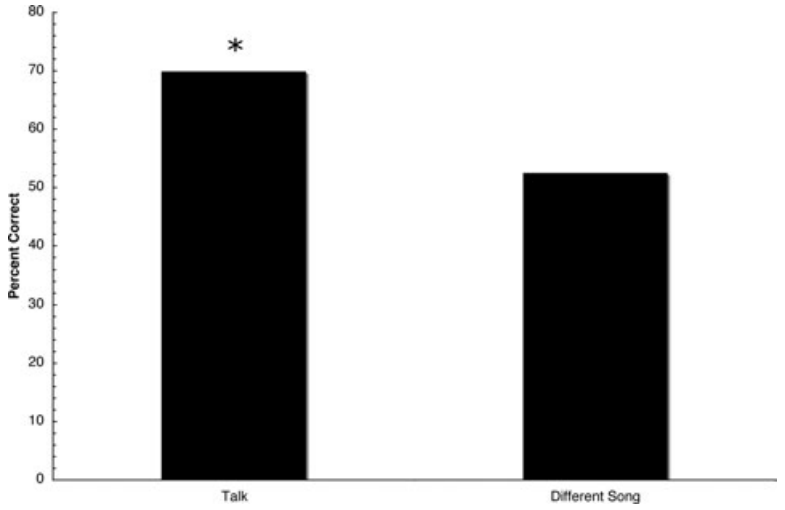

Figure 2. Adult performance on matching auditory and visual cues to identity of unfamiliar persons. *Indicates responses significantly exceeding chance levels.

would be expected by chance, $t(23) = 2.89$, $P = 0.008$.

Conclusion

These findings reveal that infants as young as 6 months of age perceive cross-modal cues to identity in infant-directed speech and singing. The relevant matching cues remain to be identified, but common temporal patterns of sound and movement are likely to be implicated. In infant-directed speech, exaggerated prosody[13] may heighten cues to identity. In infant-directed singing, elevated pitch and expressive dynamic variations[14,15] may do likewise.

Infants' ability to link auditory and visual cues to a singer's identity adds to their array of music-processing skills, which include the perception of pitch and timing relations,[16] preferential processing of consonant stimuli,[17] ease of learning foreign metrical structures,[18] and long-term memory for music.[19] Infants'

intermodal processing of music is also evident in the influence of movement on their encoding of rhythm.[20] In everyday life, cross-modal processing of music is common despite its relative absence in laboratory settings. It remains for future research to identify cross-modal parallels in expressive speech and music, including cues to identity.

Acknowledgments

Preparation of this paper was assisted by funds from the Social Sciences and Humanities Research Council of Canada.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Van Lancker, D. & J. Kreiman. 1987. Voice discrimination and recognition are separate abilities. *Neuropsychologia* **25:** 829–834.
2. Lander, K. *et al*. 2007. It's not what you say but the way you say it: matching faces and voices. *J. Exp. Psychol. Hum. Percept. Perform*. **33:** 905–914.
3. Sheffert, S.M. *et al*. 2002. Learning to recognize talkers from natural sinewave and reverse speech samples. *J. Exp. Psychol. Hum. Percept. Perform*. **28:** 1447–1469.
4. O'Toole, A.J., D.A. Roark & H. Abdi. 2002. Recognizing moving faces: a psychological and neural synthesis. *Trends Cogn. Sci.* **6:** 261–266.
5. Munhall, K.G. *et al*. 2004. Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychol. Sci.* **15:** 133–137.
6. Kamachi, M. *et al*. 2003. 'Putting the face to the voice': matching identity across modality. *Curr. Biol.* **13:** 1709–1714.
7. Bartholomeus, B. 1973. Voice identification by nursery school children. *Can. J. Psychol.* **27:** 464–472.
8. Spence, M.J., P.R. Rollins & S. Jerger. 2002. Children's recognition of cartoon voices. *J. Speech Lang. Hear. Res.* **45:** 214–222.
9. DeCasper, A.J. & W.P. Fifer. 1980. Of human bonding: newborns prefer their mothers' voices. *Science* **208:** 1174–1176.
10. Sai, F.Z. & I.W.R. Bushnell. 1988. The perception of faces in different poses by 1-month olds. *Br. J. Dev. Psychol.* **6:** 35–41.
11. Trehub, S.E. & J. Brcic. 2008. Infants hear faces and see voices. Presented at the XVIth International Conference on Infant Studies, Vancouver, Canada, March, 2008.
12. Trehub, S.E. & L.J. Trainor. 1998. Singing to infants: lullabies and play songs. *Adv. Infancy Res.* **12:** 43–77.
13. Fernald, A. 1991. Prosody in speech to children: prelinguistic and linguistic functions. *Ann. Child Dev.* **8:** 43–80.
14. Trainor, L.J. 1996. Infant preferences for infant-directed versus noninfant-directed playsongs and lullabies. *Infant Behav. Dev.* **19:** 83–92.
15. Nakata, T. & S.E. Trehub. 2009. Expressive timing and dynamics in ID and non-ID singing. *Psychomusicology* In press.
16. Trehub, S.E. & E.E. Hannon. 2006. Infant music perception: domain-general or domain-specific mechanisms? *Cognition* **100:** 73–99.
17. Trainor, L.J. & B.M. Heinmiller. 1998. The development of evaluative responses to music: infants prefer to listen to consonance over dissonance. *Infant Behav. Dev.* **21:** 77–88.
18. Hannon, E.E. & S.E. Trehub. 2005. Tuning in to musical rhythms: infants learn more readily than adults. *Proc. Natl. Acad. Sci. USA* **102:** 12639–12643.
19. Volkova, A., S.E. Trehub & E.G. Schellenberg. 2006. Infants' memory for musical performances. *Dev. Sci.* **9:** 584–590.
20. Phillips-Silver, J. & L.J. Trainor. 2005. Feeling the beat in music: movement influences rhythm perception in infants. *Science* **308:** 1430.