Current Biology 22, 113–117, January 24, 2012 ©2012 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2011.11.052

Report

Children's Development of Self-Regulation in Speech Production

Ewen N. MacDonald,[1,2,*] Elizabeth K. Johnson,[4] Jaime Forsythe,[2] Paul Plante,[2] and Kevin G. Munhall[2,3]
[1]Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads, Building 352, DK-2800 Kgs. Lyngby, Denmark
[2]Department of Psychology
[3]Department of Otolaryngology
Queen's University, Kingston, Ontario K7L3N6, Canada
[4]Department of Psychology, University of Toronto Mississauga, 3359 Mississauga Road North, Mississauga, Ontario L5L1C6, Canada

Summary

Species-specific vocalizations fall into two broad categories: those that emerge during maturation, independent of experience, and those that depend on early life interactions with conspecifics. Human language and the communication systems of a small number of other species, including songbirds, fall into this latter class of vocal learning. Self-monitoring has been assumed to play an important role in the vocal learning of speech [1–3] and studies demonstrate that perception of your own voice is crucial for both the development and lifelong maintenance of vocalizations in humans and songbirds [4–8]. Experimental modifications of auditory feedback can also change vocalizations in both humans and songbirds [9–13]. However, with the exception of large manipulations of timing [14, 15], no study to date has ever directly examined the use of auditory feedback in speech production under the age of 4. Here we use a real-time formant perturbation task [16] to compare the response of toddlers, children, and adults to altered feedback. Children and adults reacted to this manipulation by changing their vowels in a direction opposite to the perturbation. Surprisingly, toddlers' speech didn't change in response to altered feedback, suggesting that long-held assumptions regarding the role of self-perception in articulatory development need to be reconsidered.

Results

In humans, there is a clearly defined linkage between vocal tract configuration and the acoustic structure of speech. The two vocal tract configurations shown in Figure 1A have different resonant frequencies leading to the amplification of different harmonics in the speech signal. Speech researchers call these amplified harmonics "formants," and listeners rely heavily on formants to determine what consonant or vowel a speaker intended to produce. As speakers shift the configuration of their vocal tract, the formant structure of their utterances shifts accordingly. By attending to the linkage between their own unique vocal tract configurations and the resulting speech acoustics, young children could fine-tune the mapping between motor commands sent from their brains to the vocal-production organs and the resulting acoustic output produced.

In the current study, we look at real-time compensatory behavior in vowel production when auditory feedback is modified. We use a rapid signal processing system to change the formant frequencies of vowels produced by children and adults. Previous work with adults has demonstrated that when talkers receive auditory feedback in which their own vowel formants are shifted to new locations in the vowel space, they rapidly compensate for the perturbations, altering the formant frequencies of the vowels they produce in a direction opposite to the perturbation [16–19]. This response pattern has been interpreted as evidence for the existence of a predictive mechanism in speech motor control [17]. This phenomenon also demonstrates that even adult speakers remain reliant on auditory feedback to fine-tune the accuracy of their vocal productions.

We tested three different age groups of native English speakers: adults (26 adult females with a mean age of 18.9 years), young children (26 children with a mean age of 51.5 months), and toddlers (20 children with a mean age of 29.8 months). Each talker produced 50 utterances of the word "bed." To elicit these utterances from the young children and toddlers, we developed a video game in which the children would help a robot cross a virtual playground by saying the robot's "magic" word "bed" (Figure 1B). During the first 20 utterances, talkers received normal acoustic feedback through a pair of headphones. During the last 30 utterances, talkers received feedback in which the frequency of their first and second formants (F1 and F2, respectively) were perturbed using a real-time formant shifting system. F1 was increased by 200 Hz and F2 was decreased by 250 Hz. This manipulation changed talkers' productions of the word "bed" into their own voice saying the word "bad."

For each utterance, the "steady-state" F1 and F2 frequency was determined by averaging estimates of that formant from 40% to 80% of the way through the vowel. These results were then normalized for each individual by subtracting that average of that individual's baseline utterances defined as the average of the last 15 utterances before feedback was altered (i.e., utterances 6–20). For statistical analyses, individual measures of compensation in F1 and F2 were computed with the magnitude based on the difference in average frequency between the last 20 utterances (i.e., utterances 31–50) and the baseline used in normalization. The sign was determined based on whether the change in production opposed (positive) or followed (negative) the direction of the perturbation.

The normalized results, averaged across individuals in each group, are plotted in Figure 2. As in previous formant perturbation experiments [16, 19], the adults spontaneously compensated by altering the frequency of F1 and F2 in a direction opposite to that of the perturbation (top panel). The young children also compensated in a manner similar to the adults (middle panel). However, the toddlers did not alter production of F1 or F2 in response to the perturbation (bottom panel).

To verify these observations, we computed individual measures of compensation in F1 and F2. For both F1 and F2,

*Correspondence: emcd@elektro.dtu.dk

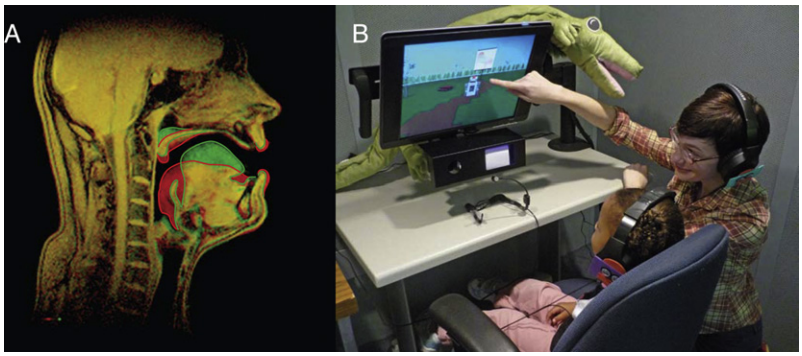

Figure 1. Articulator Positions for Different Vowels and Photo of the Experimental Setup

(A) Midsagittal adult vocal tract showing the positioning of articulators when producing two different vowels that differ in height and frontness of the vocal tract constriction. The different tongue positions result in different resonances in the vocal tract and perception of different vowels.

(B) Author J.F. explaining the computer game to a toddler.

an analysis of variance (ANOVA) revealed a significant effect between groups [F1: $F(2,69) = 7.23$, $p < 0.01$; F2: $F(2,69) = 6.38$, $p < 0.01$]. Multiple comparisons with Bonferroni correction confirmed that the compensation by the adults and young children was significantly different from that of the toddlers ($p < 0.01$ for both F1 and F2), but no significant differences between the adults and young children were observed ($p > 0.99$ for both F1 and F2).

An examination of individual's baseline utterances revealed that variability in production decreased with age. The average individual's standard deviation in F1 and F2 during production of baseline utterances is plotted in Figure 3. For both F1 and F2, an ANOVA revealed a significant effect between groups [F1: $F(2,69) = 37.23$, $p < 0.001$; F2: $F(2,69) = 22.32$, $p < 0.001$]. Multiple comparisons with Bonferroni correction confirmed that for both F1 and F2, the differences between all groups were significant ($p < 0.05$).

Discussion

Our data suggest that by the age of 4, children are monitoring their speech productions in an adult-like manner. Toddlers, in contrast, do not appear to self-regulate their vowel acoustics like adults or young children do. Feedback discrepancies with their own speech simply do not produce compensatory behaviors. At first blush, these results seem paradoxical. Perceptual attunement to the vowel space of the native language is in evidence by 6 months of age [20]. Infants readily detect small deviations in others' pronunciation of familiar words [21] and begin babbling in prosodic patterns characteristic of the language they have been exposed to [22]. By the age of 24 months, American children have an average vocabulary size of about 300 words [23]. Thus, by 2 years of age, toddlers appear to be well on their way to acquiring the sound structure of their native language. If toddlers do not automatically monitor their own speech productions for accuracy as adults and young children do, then how do they learn to produce the speech sounds used in their language community? We see two kinds of possible answers to this question: (1) explanations that are consistent with the idea that feedback error correction is important at all ages but that its role is context-dependent in young children, and (2) hypotheses that suggest that error correction based on feedback of the child's own speech develops only after the internal representation of a sound category is robust.

One context-dependent explanation for our data is that children may require different cognitive and/or social conditions to learn language at different ages. For example, Baldwin [24] showed that by 18 months, the social cue of gaze direction of a speaker is more important for infant lexical acquisition than other cues that had previously been important, such as salience of an object or temporal contiguity of object and name. Similarly, the speech processing behavior of very young children during word learning varies with different cognitive demands. For some online speech testing procedures, young children do not attach labels to objects as readily as they do if they were given more naturalistic contextual support or simpler tasks [25].

Social context might also modulate when auditory feedback can influence the sound representation. As has been shown with songbirds, social or public use of vocalizations and vocal practice in early learning can be differentiated and feedback plays a different role in each type of vocalization [26]. For our 2-year-olds, the minimal speech produced by the adults during the task may have resulted in a situation where fine-tuning of production was minimized. In addition, the words produced by the children were reinforced by the video game independent of their accuracy in producing the vowel—the robot progressed through the playground regardless of whether the child did or did not compensate. Note that this was true for both toddlers and young children and this by itself does not explain the age-related changes.

Alternatively, more in line with our second class of explanation, feedback error correction may not be adaptive during the earliest stages of word production, perhaps because of the magnitude of variability observed in the motor activities of toddlers. If production variance alone was the issue, compensations should only be observed when variability is reduced to a tolerable amount. To explore this hypothesis, we conducted two types of analyses. In the first, regressions were computed between an individual's compensation magnitude and production variability in the baseline values of F1 and F2. When the regressions were carried out within age groups and when the compensation results of the toddler and young children groups were pooled together, no significant relationship ($p > 0.3$) was found. In the second analysis, we tested whether the perturbation was influencing articulation even if the youngest children did not compensate. It is conceivable that the altered feedback might induce instability even if mature compensatory behavior was not developed. To test this, we compared the standard deviation of an individual's last 15 utterances of the baseline phase and shift phase. For both the toddlers and young children, no significant difference in standard deviation was observed for either F1 or F2. These results suggest that variability per se is not the issue.

An additional possibility in line with our second class of explanation is that the rapid growth of the vocal tract during the first two years of life may combine with motor variability to make feedback-based control suboptimal. The first couple years of life is one of the periods associated with rapid

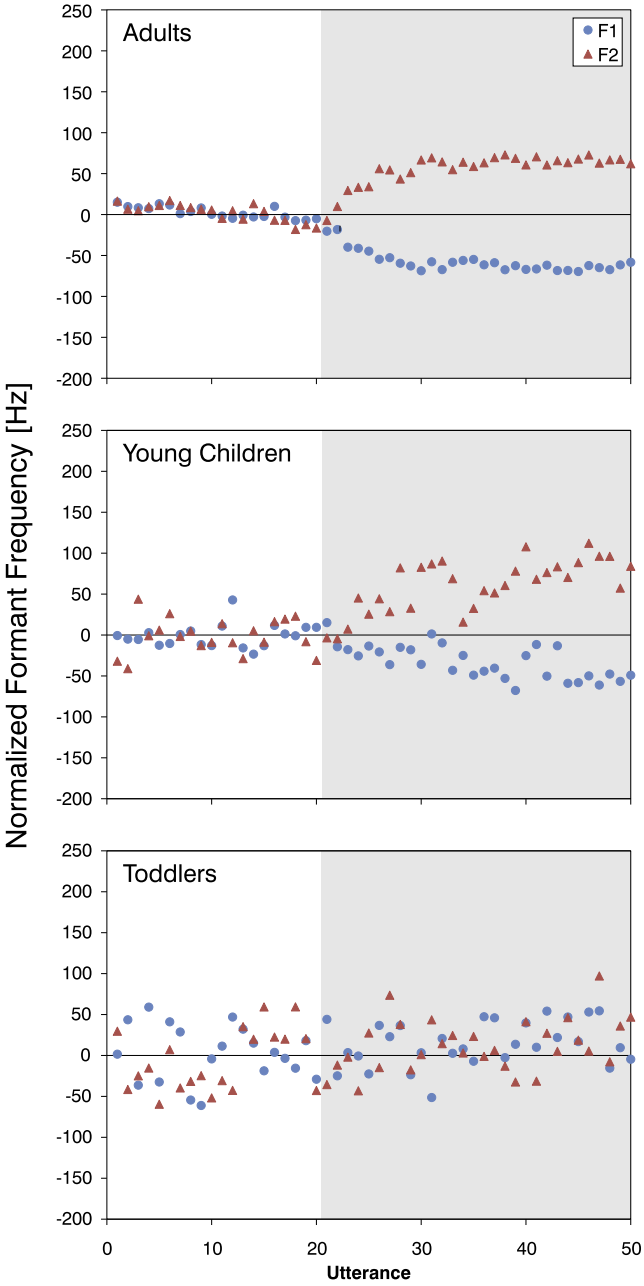

Figure 2. Normalized Formant Frequencies across Time

Normalized F1 (circles) and F2 (triangles) frequency estimates across time for adults (top panel), young children (middle panel), and toddlers (bottom panel). The shaded region indicates utterances during which talkers received altered auditory feedback.

changes in vocal tract size and configuration, primarily due to descent of the larynx [27]. A consequence of this rapid growth is abrupt change in vowel formant values between ages 1 and 4 [28]. One solution to this early phase of vocal learning is for learners to regulate their productions to the vocalizations of their communication partners rather than to use feedback from their own ill-defined targets for error correction. This suggestion is consistent with growing evidence that contingent adult behaviors shape the course of vocal learning in both birdsong acquisition and speech development [29, 30].

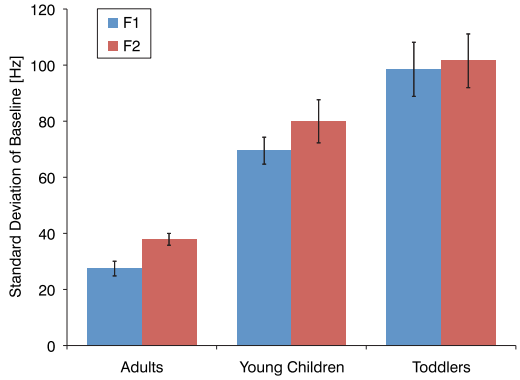

Figure 3. Variability of Normal Production

Standard deviation in F1 and F2 of an average individual's production of baseline utterances for each of the three groups. Standard error bars are shown.

The most remarkable evidence for socially guided vocal learning comes from the study of the brown-headed cowbird [31]. Juvenile males, raised in isolation with females that do not sing, nevertheless acquired mature, species-specific songs. Video analyses revealed that the immature male vocalizations were shaped by visual feedback from females (small wing strokes). Thus, without hearing mature male models, these young males were able to learn songs that contained the markers of regional dialects and songs that could strongly elicit female mating responses. As this example demonstrates, adult input to the vocal learning process can vary over a wide spectrum, ranging from an acoustic template for assessing articulation error [32, 33] to nonverbal reinforcements for correct articulation [31, 34].

The period between ages 1 and 4 is marked by many other cognitive and linguistic developments associated with speech processing. For example, there are questions concerning the immaturity of the receptive phonology of children in this age group when engaged in word learning [35], despite evidence that even younger children are able to make fine-grained speech discriminations. Although the auditory speech perception system and the auditory control of speech articulation clearly overlap and share resources, each system appears to have unique requirements and neural architectures tuned to meet those requirements. Single-cell populations in the auditory cortex of nonhuman primates have been shown to be selectively activated or inhibited during the animal's vocalizations as compared to listening to others [36, 37]. Unique functional magnetic resonance imaging activity in feedback compared to listening conditions has also been shown in humans [38]. The auditory feedback system also has different functional components, including a mapping between articulator movements and acoustics, an error detection system, and a computational model that learns from the errors and computes new trajectories for speech movements. All of these components must undergo development because the vocal tract changes in size and shape and articulatory precision changes over time. Only through the use of real-time perturbation experiments of the kind performed here will we be able to begin to tease apart the components of this complex network of processing and understand the passage to mature communication.

In summary, an age-related difference in the use of auditory feedback to control speech production was observed. When

exposed to altered feedback in which formant frequencies were perturbed, both 4-year-olds and adults compensated but 2-year-olds did not. These results suggest that either the auditory feedback component of the speech motor-control system may be suppressed in infants and toddlers or develops between 2 and 4 years of age. Although it is not possible to distinguish between these two classes of hypotheses using the present data, the finding that toddlers do not monitor their own auditory feedback in a manner similar to adults has broad implications for models of speech learning.

Experimental Procedures

Participants

The adult group consisted of 26 female undergraduate students at Queen's University (mean age of 18.9 years, range 17–22).

For the young children group, a total of 31 children between the ages of 3 and 4 years old were recruited in Kingston and Mississauga, Ontario. However, five of these children were excluded—four due to problems in tracking their formants and one due to equipment malfunction—leaving a group of 26 children with a mean age of 51.5 months and range of 43–59 months.

For the toddler group, a total of 50 2-year-old talkers were recruited in Mississauga, Ontario. Twenty-three of the toddlers did not complete the experiment. Ten of the toddlers refused to talk and 13 refused to wear the headphones. Seven of the remaining 27 toddlers that did complete the experiment were also excluded. Of these seven, six of the toddlers did not produce their utterances with a consistent timing (and thus did not receive altered feedback) and one did not produce utterances of the target vowel during the baseline phase. Twenty toddlers with a mean age of 29.8 months (range of 23–35) completed the task. When considering all the toddlers recruited for this study, one may be concerned about the high rate of attrition and the potential for selection bias. We note, however, that no compensation was observed from the toddlers included in this study. Thus, even if these toddlers were in some sense more advanced than the average 2-year-old, we can still be confident that less mature toddlers would also not compensate for the altered formant feedback.

All talkers in the experiment spoke English as their first language and reported no speech or language impairments. The protocol for this study was approved by the institutional ethics review board at both Queen's University and the University of Toronto Mississauga. All of the adult talkers provided informed consent. All of the young children and toddlers provided verbal assent and their guardians provided informed consent.

Equipment

All of the adults and 14 of the young children included in the study were tested at Queen's University. The remaining 12 young children and the 20 toddlers included in the study were tested at University of Toronto Mississauga. The same equipment was used in both locations and was identical to that reported in MacDonald et al. [19].

Talkers were seated in front of a computer monitor in a sound-insulated booth (Industrial Acoustics Company, Bronx, NY). Adult talkers were instructed to say the word "bed" at a natural rate and level when it appeared on a monitor in front of them. The young children and toddlers were instructed that they would be playing a computer game where they would help a forgetful robot move across a playground. At the beginning of each level in the game, the playground would appear with the robot at one end and a billboard with a picture of a bed at the other. The children were instructed that they could help the robot move by saying the word "bed." When a child produced an utterance of bed, an operator pressed a button and the robot advanced forward through the playground. The children were familiarized with the game using a training level that required five utterances for the robot to traverse the playground. The game was started after the training level. For each of the five levels in the game, ten utterances were required for the robot to completely traverse the playground.

The speech was recorded using a headset microphone (Shure WH20), amplified (Tucker-Davis Technologies MA3 microphone amplifier), low-pass filtered with a cutoff frequency of 4,500 Hz (Krohn-Hite 3384 filter), and digitized at 10 kHz (National Instruments PXI-8106 embedded controller). The National Instruments system generated formant estimates every nine speech samples. Based on these estimates, filter coefficients were calculated to produce formant shifts, and the filtering was conducted by the National Instruments system. To mask bone-conducted feedback, we amplified the manipulated voice signal and mixed it with speech noise (Madsen Midimate 622 audiometer) and presented over headphones (Sennheiser HD 265) such that the speech and noise were presented at approximately 80 and 50 dBA SPL respectively.

Online Formant Shifting and Detection of Voicing

Detection of voicing and formant shifting was performed as previously described by MacDonald et al. [19]. Voicing was detected using a statistical amplitude-threshold technique. The formant shifting was achieved in real-time using an infinite impulse response filter. Formants were estimated every 900 μs using an iterative Burg algorithm [39]. Filter coefficients were computed based on these estimates such that a pair of spectral zeroes was placed at the location of the existing formant frequency and a pair of spectral poles was placed at the desired frequency of the new formant.

Offline Formant Analysis

The recorded data were analyzed in the same way as that used by MacDonald et al. [19]. The boundaries of the vowel segment in each utterance were estimated using an automated process based on the harmonicity of the power spectrum. These boundaries were then inspected by hand and corrected if required.

The first three formant frequencies were estimated offline from the first 25 ms of a vowel segment using the same algorithm as that used in online shifting. The formants were estimated again after shifting the window 1 ms and repeated until the end of the vowel segment was reached. For each vowel segment, a single steady-state value for each formant was calculated by averaging the estimates for that formant from 40% to 80% of the way through the vowel. The steady-state results for F1, F2, and F3 for each individual were plotted and inspected. Any estimates that were incorrectly categorized as another (e.g., F2 being mislabeled as F1, etc.) were corrected by hand.

Acknowledgments

This research was supported by the National Institute of Deafness and Communicative Disorders Grant DC-08092 and the Natural Sciences and Engineering Research Council of Canada.

Received: October 10, 2011
Revised: November 1, 2011
Accepted: November 18, 2011
Published online: December 22, 2011

References

1. Callan, D.E., Kent, R.D., Guenther, F.H., and Vorperian, H.K. (2000). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. J. Speech Lang. Hear. Res. *43*, 721–736.
2. Howard, I.S., and Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. Mot. Contr. *15*, 85–117.
3. Bailly, G. (1997). Learning to speak. Sensorimotor control of speech movements. Speech Commun. *22*, 251–267.
4. Oller, D.K., and Eilers, R.E. (1988). The role of audition in infant babbling. Child Dev. *59*, 441–449.
5. Osberger, M.J., and McGarr, N. (1982). Speech production characteristics of the hearing impaired. In Speech and Language: Advances in Basic Research and Practice, *Volume 8*, N. Lass, ed. (New York: Academic Press), pp. 221–283.
6. Konishi, M. (1965). Effects of deafening on song development in American robins and black-headed grosbeaks. Z. Tierpsychol. *22*, 584–599.
7. Waldstein, R.S. (1990). Effects of postlingual deafness on speech production: implications for the role of auditory feedback. J. Acoust. Soc. Am. *88*, 2099–2114.
8. Nordeen, K.W., and Nordeen, E.J. (1992). Auditory feedback is necessary for the maintenance of stereotyped song in adult zebra finches. Behav. Neural Biol. *57*, 58–66.
9. Manabe, K., Sadr, E.I., and Dooling, R.J. (1998). Control of vocal intensity in budgerigars (Melopsittacus undulatus): differential reinforcement of vocal intensity and the Lombard effect. J. Acoust. Soc. Am. *103*, 1190–1198.

10. Lombard, E. (1911). Le signe de le elevation de la voix (The sign of the rise in the voice). Ann. Malad. l'Orielle Larynx Nez Pharynx (Annals of diseases of the ear, larynx, nose and pharynx) *37*, 101–119.
11. Osmanski, M.S., and Dooling, R.J. (2009). The effect of altered auditory feedback on control of vocal production in budgerigars (Melopsittacus undulatus). J. Acoust. Soc. Am. *126*, 911–919.
12. Lee, B.S. (1950). Effects of delayed speech feedback. J. Acoust. Soc. Am. *22*, 824–826.
13. Burnett, T.A., Freedland, M.B., Larson, C.R., and Hain, T.C. (1998). Voice F0 responses to manipulations in pitch feedback. J. Acoust. Soc. Am. *103*, 3153–3161.
14. Yeni-Komshian, G., Chase, R.A., and Mobley, R.L. (1968). The development of auditory feedback monitoring. II. Delayed auditory feedback studies on the speech of children between two and three years of age. J. Speech Hear. Res. *11*, 307–315.
15. Belmore, N.F., Kewley-Port, D., Mobley, R.L., and Goodman, V.E. (1973). The development of auditory feedback monitoring: delayed auditory feedback studies on the vocalizations of children aged six months to 19 months. J. Speech Hear. Res. *16*, 709–720.
16. Houde, J.F., and Jordan, M.I. (1998). Sensorimotor adaptation in speech production. Science *279*, 1213–1216.
17. Purcell, D.W., and Munhall, K.G. (2006). Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. J. Acoust. Soc. Am. *120*, 966–977.
18. Villacorta, V.M., Perkell, J.S., and Guenther, F.H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. J. Acoust. Soc. Am. *122*, 2306–2319.
19. MacDonald, E.N., Goldberg, R., and Munhall, K.G. (2010). Compensations in response to real-time formant perturbations of different magnitudes. J. Acoust. Soc. Am. *127*, 1059–1068.
20. Doupe, A.J., and Kuhl, P.K. (1999). Birdsong and human speech: common themes and mechanisms. Annu. Rev. Neurosci. *22*, 567–631.
21. Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound. Dev. Sci. *8*, 432–443.
22. de Boisson-Bardies, B. (1999). How Language Comes to Children (Cambridge: MIT press).
23. Stoel-Gammon, C. (2011). Relationships between lexical and phonological development in young children. J. Child Lang. *38*, 1–34.
24. Baldwin, D.A. (1993). Infants' ability to consult the speaker for clues to word reference. J. Child Lang. *20*, 395–418.
25. Yoshida, K.A., Fennell, C.T., Swingley, D., and Werker, J.F. (2009). Fourteen-month-old infants learn similar-sounding words. Dev. Sci. *12*, 412–418.
26. Sakata, J.T., and Brainard, M.S. (2009). Social context rapidly modulates the influence of auditory feedback on avian vocal motor control. J. Neurophysiol. *102*, 2485–2497.
27. Fitch, W.T., and Giedd, J. (1999). Morphology and development of the human vocal tract: a study using magnetic resonance imaging. J. Acoust. Soc. Am. *106*, 1511–1522.
28. Vorperian, H.K., and Kent, R.D. (2007). Vowel acoustic space development in children: a synthesis of acoustic and anatomic data. J. Speech Lang. Hear. Res. *50*, 1510–1545.
29. Goldstein, M.H., and Schwade, J.A. (2010). From birds to words: Perception of structure in social interaction guides vocal development and language learning. In The Oxford Handbook of Developmental and Comparative Neuroscience, M.S. Blumberg, J.H. Freeman, and S.R. Robinson, eds. (Oxford: Oxford University Press), pp. 708–729.
30. Kuhl, P.K., Tsao, F.-M., and Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. Proc. Natl. Acad. Sci. USA *100*, 9096–9101.
31. West, M.J., and King, A.P. (1988). Female visual displays affect the development of male song in the cowbird. Nature *334*, 244–246.
32. Kuhl, P.K., and Meltzoff, A.N. (1996). Infant vocalizations in response to speech: vocal imitation and developmental change. J. Acoust. Soc. Am. *100*, 2425–2438.
33. Papousek, M., and Papousek, H. (1989). Forms and functions of vocal matching in interactions between mothers and their precanonical infants. First Lang. *9*, 137–158.
34. Goldstein, M.H., King, A.P., and West, M.J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. Proc. Natl. Acad. Sci. USA *100*, 8030–8035.
35. Stager, C.L., and Werker, J.F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. Nature *388*, 381–382.
36. Eliades, S.J., and Wang, X. (2003). Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations. J. Neurophysiol. *89*, 2194–2207.
37. Eliades, S.J., and Wang, X. (2008). Neural substrates of vocalization feedback monitoring in primate auditory cortex. Nature *453*, 1102–1106.
38. Zheng, Z.Z., Munhall, K.G., and Johnsrude, I.S. (2010). Functional overlap between regions involved in speech perception and in monitoring one's own voice during speech production. J. Cogn. Neurosci. *22*, 1770–1781.
39. Orfanidis, S.J. (1988). Optimum Signal Processing, An Introduction (New York: MacMillan), pp. 590.