UC Merced
UC Merced Previously Published Works

Title
Phonetic matching of auditory and visual speech develops during childhood: Evidence from sine-wave speech

Permalink
https://escholarship.org/uc/item/31k624nt

Authors
Baart, Martijn
Bortfeld, Heather
Vroomen, Jean

Publication Date
2015

DOI
10.1016/j.jecp.2014.08.002

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Journal of Experimental Child Psychology 129 (2015) 157–164

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

Brief Report

Phonetic matching of auditory and visual speech develops during childhood: Evidence from sine-wave speech

Martijn Baart [a,*], Heather Bortfeld [b,c], Jean Vroomen [d]

[a] BCBL. Basque Center on Cognition, Brain, and Language, 20009 Donostia (San Sebastián), Spain
[b] Department of Psychology, University of Connecticut, Storrs, CT 06269, USA
[c] Haskins Laboratories, New Haven, CT 06511, USA
[d] Department of Cognitive Neuropsychology, Tilburg University, 5000 LE Tilburg, The Netherlands

A R T I C L E   I N F O

Article history:
Received 16 May 2014
Revised 12 August 2014
Available online 23 September 2014

Keywords:
Audiovisual speech
Cross-modal correspondence
Phonetic cues
Temporal cues
Sine-wave speech
Development

A B S T R A C T

The correspondence between auditory speech and lip-read information can be detected based on a combination of temporal and phonetic cross-modal cues. Here, we determined the point in developmental time at which children start to effectively use phonetic information to match a speech sound with one of two articulating faces. We presented 4- to 11-year-olds ($N = 77$) with three-syllabic sine-wave speech replicas of two pseudo-words that were perceived as non-speech and asked them to match the sounds with the corresponding lip-read video. At first, children had no phonetic knowledge about the sounds, and matching was thus based on the temporal cues that are fully retained in sine-wave speech. Next, we trained all children to perceive the phonetic identity of the sine-wave speech and repeated the audiovisual (AV) matching task. Only at around 6.5 years of age did the benefit of having phonetic knowledge about the stimuli become apparent, thereby indicating that AV matching based on phonetic cues presumably develops more slowly than AV matching based on temporal cues.

© 2014 Elsevier Inc. All rights reserved.

* Corresponding author.
E-mail address: m.baart@bcbl.eu (M. Baart).

http://dx.doi.org/10.1016/j.jecp.2014.08.002
0022-0965/© 2014 Elsevier Inc. All rights reserved.

Introduction

Although human infants are sensitive to audiovisual (AV) phonetic congruence in speech (e.g., Burnham & Dodd, 1996; Kuhl & Meltzoff, 1982; Patterson & Werker, 2003), the ability to extract phonetic content from visual speech improves dramatically during childhood and into puberty (e.g., Desjardins, Rogers, & Werker, 1997; Erdener & Burnham, 2013; Hockley & Polka, 1994; Kushnerenko, Teinonen, Volein, & Csibra, 2008; Massaro, 1984; McGurk & MacDonald, 1976; Ross et al., 2011; Sekiyama & Burnham, 2008). Although this may possibly be explained by a U-shaped trajectory of AV speech development (see, e.g., Knowland, Mercure, Karmiloff-Smith, Dick, & Thomas, 2014, for a similar argument), infants' use of phonetic information is not mandatory (Desjardins & Werker, 2004).

Recently, Baart, Vroomen, Shaw, and Bortfeld (2014) argued that infants might not need phonetic information to detect correspondence in AV speech whenever salient non-phonetic cues are available. They compared adults and infants on AV matching of three-syllable strings with one of two simultaneously delivered lip-read videos. The speech sounds were either natural speech or artificial sine-wave speech (Remez, Rubin, Pisoni, & Carrell, 1981). Critically, the temporal dynamics of natural speech are retained in sine-wave speech, and this information was thus available to all listeners. AV correspondence detection was 25% higher for adults who heard natural speech than for those who heard sine-wave speech, which shows that phonetic knowledge was beneficial to them. However, adults who heard sine-wave speech did match the sound with the lip-read information significantly above chance, presumably because they detected the temporal AV correspondence. In contrast, infants did not seem to benefit from the phonetic information given that their above-chance performance was alike for natural speech and sine-wave speech, which led to the conclusion that infants had presumably relied only on the temporal AV cues. If so, it is conceivable that children would also be able to rely on temporal cues because sensitivity to AV synchrony increases during development (e.g., Grant, van Wassenhove, & Poeppel, 2004; Lewkowicz, 2010). In the same vein, van Linden and Vroomen (2008) showed that whereas 8-year-olds learn to categorize ambiguous speech based on previously seen lip-read information, 5-year-olds do not. This supports the notion that somewhere in between 5 and 8 years of age, phonetic information in the AV speech signal becomes beneficial.

Here, we directly assessed this hypothesis by testing 4- to 11-year-olds on their ability to match a sine-wave speech token with one of two simultaneously presented lip-read speech videos. The elegance of sine-wave speech is that listeners can be tested in a perceptual non-speech mode and/or a perceptual speech mode. In the first mode, listeners do not have access to the phonetic auditory content; in the second, they do. Once listeners are in speech mode, they cannot switch back to the non-speech mode. Therefore, a within-participant design requires the speech mode test to be preceded by the non-speech mode test (see, e.g., Tuomainen, Andersen, Tiippana, & Sams, 2005). Thus, we first established children's AV matching capacity while participants were in non-speech mode, assuming that they could rely only on temporal cues to detect AV correspondence. The critical manipulation consisted of subsequent training in which children were informed about the speech-like nature of the sine-wave tokens so that they perceived the phonetic identity of the sounds (children were now in speech mode, which presumably affects AV integration based on phoneme-to-viseme mapping), after which we again measured AV matching. The difference in performance on each task (the "speech mode effect") was interpreted as a perceptual benefit of phonetic information in detecting AV speech correspondence. In keeping with the literature (e.g., van Linden & Vroomen, 2008), we expected this benefit to become apparent between 5 and 8 years of age and to further increase with age.

Method

Participants

A total of 77 Dutch children between 4 and 11 years of age with normal hearing and normal or corrected-to-normal vision participated in the experiment. Children were divided into three groups according to elementary school grade. In the youngest group ($n = 23$), the ages ranged between 4.2

and 6.8 years (mean = 5.6). The age range in the second group ($n = 27$) was between 7.3 and 9.3 years (mean = 8.0), and in the oldest group ($n = 27$) the ages ranged between 9.2 and 11.4 years (mean = 10.0). The 5.6-year-old group (hereafter, the mean ages are used as group labels) was recruited from the elementary school "De Peppel" in Dussen, and all other children attended the "Eerste Montessorischool" in Bergen op Zoom (both schools are located in the same province of The Netherlands). Parental consent was obtained (through an opt-out system) prior to testing. Four children were considered as outliers and were excluded from analyses (see Results for details).

Stimuli

Stimulus materials were the same as used in Baart, Vroomen, and colleagues (2014). The audio of two AV recordings of a female Dutch speaker producing the three-syllable pseudo-words "kalisu" and "mufapi" was transformed into three-tone sine-wave speech by replacing the first three formants with sinusoids that tracked the formants' center frequencies. Videos of the lip-read speech were temporally aligned with the audio relative to the onset of the initial syllable and total duration (46 frames, ~1535 ms).

Procedure

Visual stimuli were presented on a laptop (17-inch Dell Latitude E5500, 60-Hz vertical refresh rate). Sounds were delivered at a comfortable listening level through two external speakers placed to the left and right of the screen. Total testing lasted approximately 15 min and was composed of four phases: non-speech mode training, a non-speech mode AV matching task, speech mode training, and a speech mode AV matching task.

Non-speech mode training

Children got acquainted with the sine-wave stimuli by hearing them in alternating order (six presentations per stimulus) while a written number (i.e., "sound 1" for "kalisu" and "sound 2" for "mufapi") appeared on the screen. The experimenter also read out the labels before the sounds were delivered. Children then labeled 12 sine-wave tokens (6 per stimulus, delivered in random order) as "1" or "2" through a verbal response that was keyed in by the experimenter on the laptop's keyboard.

Non-speech mode AV matching

As in Baart, Vroomen, and colleagues (2014), the two videos with lip-read speech were displayed side-by-side while one of the two corresponding sine-wave speech stimuli was played. There were four different conditions based on counterbalancing sound identity ("kalisu" or "mufapi") and the side of the video (left or right) that matched the sound. These four conditions were presented 12 times each, yielding 48 trials. For each trial, children were asked to indicate whether the sound they heard matched the left or right screen. Importantly, no reference was made to the speech-like nature of the sine-wave speech. Indeed, none of the children perceived the sounds as speech, as assessed by questions immediately after this AV matching task.

Speech mode training

Next, children were informed about the speech-like nature of the stimuli. They then underwent a short training period during which each of the sine-wave tokens was preceded by its natural speech version ("kalisu" or "mufapi") and was accompanied by an alphabetic representation ("kalisu" or "mufapi") on the screen. Each of the natural speech–sine-wave speech pairs was played six times in alternating order. After this training, both sounds were presented six times in random order and children were asked to label the sounds as "kalisu" and "mufapi" instead of as "1" and "2".

Speech mode AV matching

The matching task and procedures were the same as before (see "Non-speech mode AV matching" section above), with the only difference being that children were now informed about the phonetic nature of the sine-wave tokens.

Table 1
Mean proportions of correct auditory training responses and correct AV matches for non-speech mode and speech mode and the difference between both modes for the three different groups.

Mean age (years)	Group-averaged proportion							
	Correct auditory training responses				Correct AV matching responses			
	Overall	NSM	SM	Difference	Overall	NSM	SM	Difference
5.6	.69 (.18)	.64 (.20)	.74 (.29)	.10 (.35)	.58 (.15)	.61 (.16)	.55 (.19)	−.06 (.18)
8.0	.79 (.22)	.86 (.17)	.73 (.33)	−.13 (.30)	.66 (.18)	.60 (.17)	.71 (.23)	.11 (.17)
10.0	.84 (.18)	.86 (.20)	.81 (.34)	−.05 (.43)	.74 (.19)	.68 (.20)	.81 (.22)	.13 (.19)
β	.07	.11	.03		.08	.03	.10	

Note. Standard deviations are in parentheses. β indicates the linear trend coefficient of performance across groups. NSM is non-speech mode; SM is speech mode.

Results

We computed the proportion of correct sound identification responses during both trainings and the proportion of correct AV matches during both matching tasks. Four children were excluded from the analyses because their performance on one or more of the tasks was outside of a ±2.5-standard deviation range from the group average for that particular task; three children were from the 10.0-year-old group (one had low performance in AV matching in non-speech mode and two had low performance in non-speech mode training), and one child was from the 8.0-year-old group (low performance in non-speech mode training). The group averages for the remaining 73 participants are provided in Table 1.

A 2 (Stimulus Identity: kalisu or mufapi) × 2 (Mode: non-speech or speech) × 3 (Group: 5.6-, 8.0-, or 10.0-year-olds) mixed-effects repeated-measures analysis of variance (ANOVA) on the proportion of correct training responses produced a main effect of group, $F(2,70) = 3.52$, $p = .03$, $\eta_p^2 = .09$, because overall training performance was lower for the 5.6-year-old group than for the 10.0-year-old group, $t(45) = 2.76$, $p < .01$, $d = 0.82$ (see also Table 1). The other between-group comparisons did not reach significance (*p*s > .05). The ANOVA produced no significant main effect of stimulus identity or mode, and there were no significant interactions between (any combination of) factors (*p*s > .08). The average proportions of correct training responses were .79 for non-speech mode and .76 for speech mode.

Next, we performed an ANOVA on the proportion of correct AV matches with the same factors (see Table 1). This ANOVA revealed a main effect of group $F(2,70) = 4.99$, $p < .01$, $\eta_p^2 = .12$, because the proportion of correct matches was larger for the 10.0-year-old group than for the 5.6-year-old group, $t(45) = 3.23$, $p < .01$, $d = 0.96$ (the other two between-group comparisons yielded *p*s > .05). There was also a main effect of mode, $F(1,70) = 8.44$, $p < .01$, $\eta_p^2 = .11$, because the average proportion of correct matches was approximately 7% higher in speech mode than in non-speech mode. Critically, there was an interaction between group and mode, $F(2,70) = 8.11$, $p < .01$, $\eta_p^2 = .19$, because the proportion of correct AV matches in speech mode was higher than that in non-speech mode for both the 8.0- and 10.0-year-old groups, $t(25) = 3.24$, $p < .01$, $d = 0.55$ and $t(23) = 3.52$, $p < .01$, $d = 0.64$, respectively (see also Table 1), but not for the 5.6-year-old group ($p = .13$).[1]

In Fig. 1, we plotted performance on the AV matching tasks as a function of age rather than school grade. There was a significant positive correlation, $r(71) = .44$, $p < .01$, between age and AV matching when in speech mode (see Fig. 1A), but the correlation was not significant when the sine-wave speech was perceived as non-speech, $r(71) = .21$, $p = .08$. This was further underscored by the correlation between age and the speech mode effect, $r(71) = .34$, $p < .01$, which was calculated by subtracting the proportion of correct AV matches in non-speech mode from speech mode (see Fig. 1B).

[1] A pilot study with adults ($n = 6$) revealed a .18 increase from non-speech mode to speech mode, which is in between the 10.0-year-old group and the .25 effect when non-speech mode was compared with natural speech (Baart, Vroomen, et al., 2014).

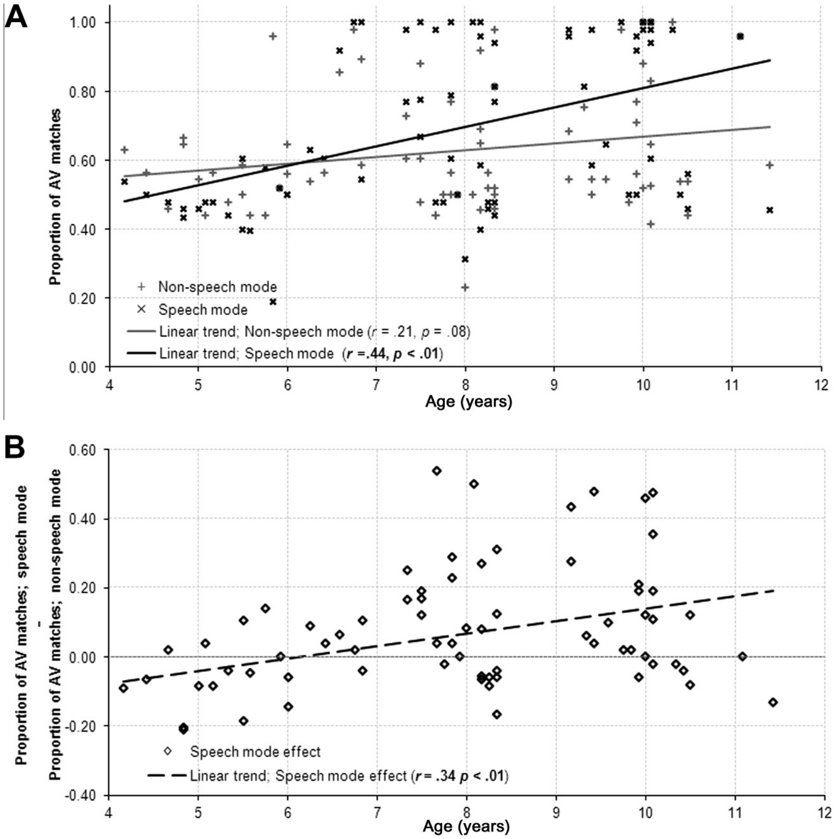

Fig. 1. (A) Scatter plot of age and proportion of correct AV matches when children were in non-speech and speech modes as well as the linear trends. (B) Scatter plot of age and the speech mode effect.

Discussion

We examined the age at which children can use phonetic information to match sine-wave speech with lip-read information. Children (4–11 years of age) were tested twice in an AV matching task. In the first test they were naive to the speech-like nature of the sounds (they were in non-speech mode), and in the second test they were informed that the sine-wave tokens were derived from natural speech (they were in speech mode). Results showed that the two groups of older children performed better in AV matching when in speech mode, whereas for the youngest group there was no such benefit. This pattern was predicted and is in line with the notion that the ability to extract phonetic content from lip-read speech develops during childhood. More specifically, Fig. 1B indicates that at around 6.5 years of age the development of phonetic processing reaches a critical point at which it becomes beneficial for AV speech perception—the point at which AV matching improved when children were made aware of the phonetic content of the sounds by being put into speech mode.

In a previous study where preverbal infants' matching of sine-wave speech with lip-read speech was tested (Baart, Vroomen, et al., 2014), it could not be established whether infants were in speech mode or not. In contrast, here we explicitly asked children whether they had perceived the sounds as speech after the first test, and we found no evidence for that. This suggests that all children can rely on non-phonetic cross-modal cues (most likely temporal) to match artificial speech sounds to an articulating face without being aware of the phonetic content. As described in Baart, Vroomen, and colleagues (2014), the sound of the second syllable was asynchronous (∼200 ms) with the incongruent lip-read video. Even though there is no behavioral evidence that infants can detect this asynchrony (e.g., Kopp, 2014; Lewkowicz, 2010), the 6-month-old infant brain is sensitive to a 200-ms offset between the unimodal signals (Kopp, 2014). Lewkowicz (2010) had proposed that the infant system may be biased toward the correlation between the auditory and visual speech signals as it exists in natural situations. If so, it seems likely that the children we tested could also rely on the temporal correlation to detect the AV correspondence (note that adults may infer a causal relationship between sight and sound even when the two are asynchronous; Parise, Spence, & Ernst, 2012).

Of relevance are studies that used sine-wave speech in behavioral and electrophysiological measures to demonstrate that different properties of the AV speech signal (e.g., temporal features vs. phonetic content) are integrated at different levels in the processing chain (Baart, Stekelenburg, & Vroomen, 2014; Baart, Vroomen, et al., 2014; Eskelund, Tuomainen, & Andersen, 2011; Stekelenburg & Vroomen, 2012; Tuomainen et al., 2005; Vroomen & Baart, 2009; Vroomen & Stekelenburg, 2011). The AV matching paradigm used in the current study indicates that it is likely that such a staged process also occurs in children; children showed a "top-up" benefit (above and beyond their already above-chance performance in non-speech mode) from having phonetic knowledge about the stimuli, but only after approximately 6.5 years of age, indicating that sufficient accrual of phonetic knowledge had occurred by then to influence the AV matching of the degraded stimuli.

As mentioned, there is a well-documented developmental trajectory for when lip-read speech influences children's auditory speech perception, with changes that extend into adulthood (e.g., Hockley & Polka, 1994; McGurk & MacDonald, 1976; Ross et al., 2011). The current findings clearly align with previous work on developmentally mediated changes in AV integration. Moreover, a recent electrophysiological study determined the neural underpinnings related to phonetic processing in children (Knowland et al., 2014) based on the fact that in adults the auditory N1 and P2 components are modulated in amplitude and latency by lip-read speech (e.g., van Wassenhove, Grant, & Poeppel, 2005). The findings from children demonstrated that the relative difference in P2 amplitude between auditory and AV speech increased between 6 and 12 years of age (Knowland et al., 2014). Given that the P2 modulations induced by lip-read speech reflect a phonetic stage of processing (as demonstrated with sine-wave speech; see Baart, Stekelenburg, et al., 2014), it seems that the changes in the evoked P2 response from 6 to 12 years of age, as observed by Knowland and colleagues (2014), are tied to ongoing development of phonetic processing. Data from the current study further corroborate this. Even still, P2 responses from the 12-year-olds indicated remaining immaturity in that they were not sensitive to AV phonetic incongruency (Knowland et al., 2014). This is in contrast to adults, for whom the P2 is quite sensitive to phonetic congruency (Klucharev, Möttönen, & Sams, 2003).

Interestingly, the infant brain is also sensitive to phonetic information in AV speech (Bristow et al., 2009; Kushnerenko et al., 2008). For instance, 6- to 9-month-olds show a lip-read-induced reduction in P2 amplitude in response to AV congruent stimuli (which hints at lip-read-induced facilitation), and their mismatch response to incongruent stimuli (A/b/V/g) is smaller for those infants who look longer at the mouth during stimulation, possibly because longer looking times are related to enhanced use of lip-read information that facilitates perceptual union of the unimodal inputs (Kushnerenko, Tomalski, Ballieux, Potton, et al., 2013; Kushnerenko, Tomalski, Ballieux, Ribeiro, et al., 2013).

As alluded to in the Introduction, the use of phonetic information may follow a U-shaped developmental course and the transition period in childhood (i.e., the plateau in the U-shaped trajectory; Smith & Thelen, 2003) may be preceded by early sensitivity and followed by later maturation (see Jerger, Damian, Spence, Tye-Murray, & Abdi, 2009, for indirect evidence where AV speech distractors were shown to affect picture naming in 4-year-olds and 10- to 14-year-olds but not in 5- to 9-year-olds). According to this view, the early signs of phonetic congruency processing in the infant brain may thus reflect an early sensitivity, which is followed by a transition during childhood when processing of phonetic congruence matures toward a stable adult state.

Another reason why children may become increasingly sensitive to lip-read speech as they mature is the onset and development of reading. Lip-reading abilities are related to reading abilities (de Gelder & Vroomen, 1998), and reading skills predict children's (language-specific) speech perception (Burnham, 2003), possibly because relatively high reading and lip-reading abilities are indicators of a stronger native language bias (Erdener & Burnham, 2013). Specifically, reading may modulate perceptual attunement to the native language, which in turn modulates AV speech integration (Erdener & Burnham, 2013),[2] which itself varies as a function of the nature of the native language (e.g., AV integration increases between 6 and 8 years of age for English children but not for Japanese children; Sekiyama & Burnham, 2008).

Taken together, there is much evidence in support of continual development of phonetic processing from childhood into adulthood. Here, we showed that after approximately 6.5 years of age children can effectively use phonetic cues to match a speech sound with the corresponding lip movements. More generally, we demonstrated that sine-wave speech provides an effective tool that can be used within participants to investigate the development of AV speech perception, opening up a variety of possibilities for future work with additional (e.g., neurophysiological) measures.

Conclusions

We used sine-wave speech as a tool to investigate the developmental trajectory underlying AV speech perception. We observed that children started using phonetic information above and beyond the non-phonetic (temporal) correlation between audio and visual speech only at around 6.5 years of age.

Acknowledgment

This research was supported by National Institutes of Health Grant R01 DC010075 to Heather Bortfeld.

References

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia, 53*, 115–121.
Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition, 130*, 31–43.
Bristow, D., Dehaene-Lamberts, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., et al (2009). Hearing faces: How the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience, 21*, 905–921.
Burnham, D. (2003). Language specific speech perception and the onset of reading. *Reading and Writing, 16*, 573–609.
Burnham, D., & Dodd, B. (1996). Auditory–visual speech perception as a direct process: The McGurk effect in human infants and across languages. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines* (pp. 103–114). Berlin: Springer-Verlag.
de Gelder, B., & Vroomen, J. (1998). Impaired speech perception in poor readers: Evidence from hearing and speech reading. *Brain and Language, 64*, 269–281.
Desjardins, R. N., Rogers, J., & Werker, J. F. (1997). An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks. *Journal of Experimental Child Psychology, 66*, 85–110.
Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology, 45*, 187–203.
Erdener, D., & Burnham, D. (2013). The relationship between auditory–visual speech perception and language-specific speech perception at the onset of reading instruction in English-speaking children. *Journal of Experimental Child Psychology, 116*, 120–138.
Eskelund, K., Tuomainen, J., & Andersen, T. S. (2011). Multistage audiovisual integration of speech: Dissociating identification and detection. *Experimental Brain Research, 208*, 447–457.
Grant, K. W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Communication, 44*, 43–53.
Hockley, N. S., & Polka, L. A. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm. *Journal of the Acoustical Society of America, 96*, 3309.
Jerger, S., Damian, M. F., Spence, M. J., Tye-Murray, N., & Abdi, H. (2009). Developmental shifts in children's sensitivity to visual speech: A new multimodal picture–word task. *Journal of Experimental Child Psychology, 102*, 40–59.
Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research, 18*, 65–75.
Knowland, V. C., Mercure, E., Karmiloff-Smith, A., Dick, F., & Thomas, M. S. (2014). Audio–visual speech perception: A developmental ERP investigation. *Developmental Science, 17*, 110–124.
Kopp, F. (2014). Audiovisual temporal fusion in 6-month-old infants. *Developmental Cognitive Neuroscience, 9*, 56–67.
Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138–1141.

[2] We obtained post hoc reading scores for all but one child in the two older groups. These were indeed positively correlated with the proportion of correct training responses and AV matches in speech mode ($ps < .03$) but not in non-speech mode ($ps > .07$).

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 11442–11445.

Kushnerenko, E., Tomalski, P., Ballieux, H., Potton, A., Birtles, D., Frostick, C., et al (2013). Brain responses and looking behavior during audiovisual speech integration in infants predict auditory speech comprehension in the second year of life. *Frontiers in Psychology, 4*, 432.

Kushnerenko, E., Tomalski, P., Ballieux, H., Ribeiro, H., Potton, A., Axelsson, E. L., et al (2013). Brain responses to audiovisual speech mismatch in infants are associated with individual differences in looking behaviour. *European Journal of Neuroscience, 8*, 3363–3369.

Lewkowicz, D. J. (2010). Infant perception of audio–visual speech synchrony. *Developmental Psychology, 46*, 66–77.

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development, 55*, 1777–1788.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology, 22*, 46–49.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*, 191–196.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947–949.

Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience, 33*, 2329–2337.

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory–visual speech perception. *Developmental Science, 11*, 306–320.

Smith, L., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences, 7*, 343–348.

Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological evidence for a multisensory speech-specific mode of perception. *Neuropsychologia, 50*, 1425–1431.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio–visual speech perception is special. *Cognition, 96*, B13–B22.

van Linden, S., & Vroomen, J. (2008). Audiovisual speech recalibration in children. *Journal of Child Language, 35*, 809–822.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 1181–1186.

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition, 110*, 254–259.

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition, 118*, 75–83.