Title

Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants

Journal

Cognition, 130(1)

ISSN

0010-0277

Authors

Baart, Martijn
Vroomen, Jean
Shaw, Kathleen
et al.

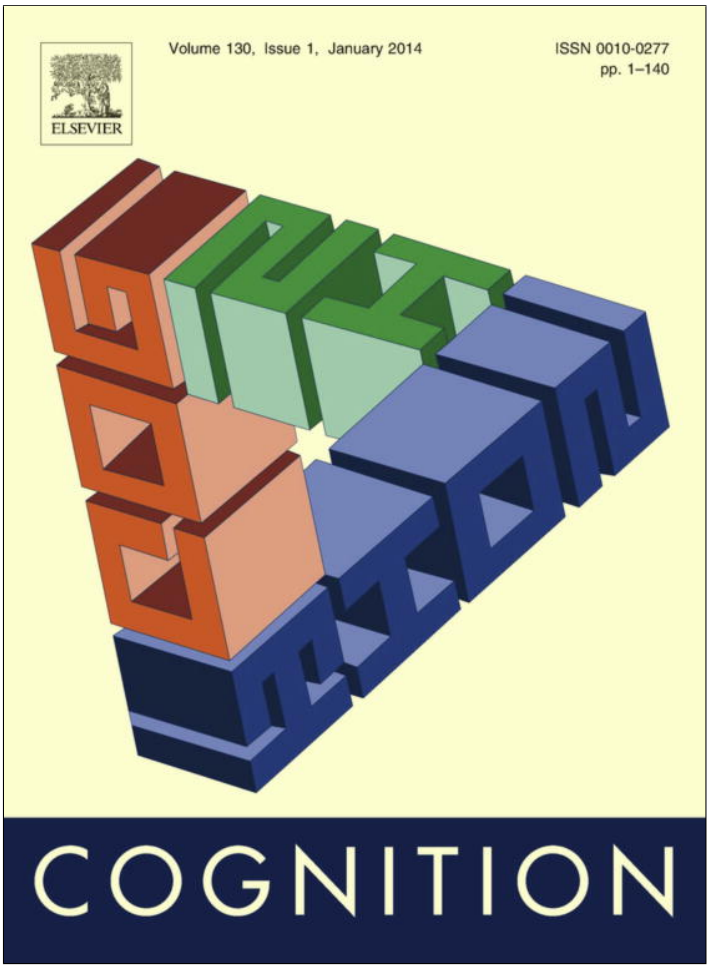
Publication Date

2014

DOI

10.1016/j.cognition.2013.09.006

Peer reviewed



Cognition 130 (2014) 31–43

Contents lists available at ScienceDirect

Cognition

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

Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants

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

[a] Basque Center on Cognition, Brain and Language, Donostia, Spain
[b] Dept. of Psychology, Tilburg University, Tilburg, The Netherlands
[c] Dept. of Psychology, University of Connecticut, Storrs, CT, United States
[d] Haskins Laboratories, New Haven, CT, United States

ARTICLE INFO

Article history:
Received 26 November 2012
Revised 24 August 2013
Accepted 20 September 2013
Available online 18 October 2013

Keywords:
Audiovisual speech integration
Adults
Infants
Phonetic correspondence
Sine-wave speech

ABSTRACT

Infants and adults are well able to match auditory and visual speech, but the cues on which they rely (viz. temporal, phonetic and energetic correspondence in the auditory and visual speech streams) may differ. Here we assessed the relative contribution of the different cues using sine-wave speech (SWS). Adults ($N = 52$) and infants ($N = 34$, age ranged in between 5 and 15 months) matched 2 trisyllabic speech sounds ('kalisu' and 'mufapi'), either natural or SWS, with visual speech information. On each trial, adults saw two articulating faces and matched a sound to one of these, while infants were presented the same stimuli in a preferential looking paradigm. Adults' performance was almost flawless with natural speech, but was significantly less accurate with SWS. In contrast, infants matched the sound to the articulating face equally well for natural speech and SWS. These results suggest that infants rely to a lesser extent on phonetic cues than adults do to match audio to visual speech. This is in line with the notion that the ability to extract phonetic information from the visual signal increases during development, and suggests that phonetic knowledge might not be the basis for early audiovisual correspondence detection in speech.

1. Introduction

Human speech is inherently audiovisual (AV) as the auditory speech signal is accompanied by the articulating mouth of a speaker (here referred to as 'visual speech'). It has been demonstrated repeatedly that both the adult and infant brain combine auditory and visual speech into a single percept (e.g., Burnham & Dodd, 2004; Burnham & Sekiyama, 2004; Kuhl & Meltzoff, 1982, 1984; McGurk & MacDonald, 1976; Patterson & Werker, 1999, 2003; Rosenblum, Schmuckler, & Johnson, 1997; Sumby & Pollack, 1954). In order to do this, listeners may rely on at least three cross-modal cues in the signal: (i) temporal cues, (ii) energetic cues, and (iii) phonetic cues.

Cross-modal temporal cues consist of bimodally shared characteristics such as a speaker's speech rate and the AV onset of syllables. When unimodal signals are presented out of synchrony, both adults and infants (4 month-olds and older and given sufficiently offset unimodal stimuli) are able to detect the AV asynchrony (e.g., Grant, van Wassenhove, & Poeppel, 2004; Lewkowicz, 2000, 2010; van Wassenhove, Grant, & Poeppel, 2007; Vatakis & Spence, 2006), indicating that both adults and infants are sensitive to violations in the AV temporal correlation. In fact, as demonstrated with non-speech stimuli, listeners may even rely on this correlation to infer causal relationships (Parise, Spence, & Ernst, 2012).

* Corresponding author. Address: Department of Psychology, University of Connecticut, 406 Babbidge Road, Unit 1020, Office: BOUS 151, Storrs, CT 06269-1020, United States. Tel.: +1 860 4860919; fax: +1 860 4862760.
E-mail address: heather.bortfeld@uconn.edu (H. Bortfeld).

0010-0277/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2013.09.006

Energetic cues in the AV signal can be defined as the correlation between acoustic energy and the visible articulators (Grant, 2001; Grant & Seitz, 2000; Grant et al., 2004). The energy in human speech mainly stems from the (invisible) vocal folds. Nevertheless, there is (a modest) cross-modal correlation between the visible movements of the lips (e.g., inter-lip distance or area of mouth) and the acoustic speech envelope because, in general, there is more acoustic energy when the mouth is open rather than closed.

The third cross-modal cue in the AV speech signal is related to the phonetic correspondence between the visual signal and the sound (e.g., a listener recognizes that a bilabial closure is specific to speech and corresponds to /m/ or /p/, but not to /k/ or /s/). Although even infants are sensitive to phonetic information in the (AV) speech signal (Burnham & Dodd, 1996, 2004; Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Jusczyk & Luce, 1994; Kuhl et al., 2006; Kushnerenko, Teinonen, Volein, & Csibra, 2008; Rosenblum et al., 1997; Swingley, Pinto, & Fernald, 1999), the ability to extract phonetic content from visual speech increases with age and develops well beyond puberty (Bruce et al., 2000; Desjardins, Rogers, & Werker, 1997; Hockley & Polka, 1994; Massaro, 1984; McGurk & MacDonald, 1976; Ross et al., 2011; Sekiyama & Burnham, 2004). Of course, phonetic AV correspondence is closely linked to AV energetic cues as the specific shape of the vocal filter determines both the energetic correlation and the phonetic quality of specific speech tokens (e.g., Fant, 1960; Lieberman & Blumstein, 1988).

Separating energetic cues from phonetic cues may thus be rather challenging, but it is not entirely impossible. For instance, the speech signal can be transformed into so-called sine-wave speech (SWS, see Remez, Rubin, Pisoni, & Carrell, 1981) in which the center-frequencies of the first three formants are replaced by sinusoids. In SWS, the temporal properties of natural speech are completely retained, as well as critical energetic properties in F2 and F3 (Grant & Seitz, 2000). In contrast, the detail of phonetic information is severely compromised in SWS such that listeners typically do not perceive SWS as speech but rather as non-speech whistles or bleeps (e.g., Remez et al., 1981) and need explicit training to perceive the exact phonetic content of SWS (e.g., Eskelund, Tuomainen, & Andersen, 2011; Tuomainen, Andersen, Tiippana, & Sams, 2005; Vroomen & Baart, 2009; Vroomen & Stekelenburg, 2011).

Recent studies using SWS have provided converging evidence in support of the notion that AV speech integration in adults is achieved at multiple levels (Schwartz, Berthommier, & Savariaux, 2004). For example, visual speech-induced phonetic biases in auditory speech identification occur only for listeners who are trained to hear the phonetic content in the SWS sounds (Eskelund et al., 2011; Tuomainen et al., 2005; Vroomen & Baart, 2009; Vroomen & Stekelenburg, 2011), whereas perception of AV synchrony (Vroomen & Stekelenburg, 2011) and a visual speech-induced detection advantage for SWS embedded in noise (Eskelund et al., 2011) are independent of participants' phonetic interpretation of the SWS stimuli. But the developmental timeline underlying this multilevel integration process is unclear.

It is well established that, when presented with two simultaneous videos of a speaker articulating a single vowel, infants prefer to look at the speaker whose visual speech matches the speech sound they are hearing (e.g., Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2003). For example, Kuhl and Meltzoff (1982) presented 18- to 20-week-old infants with two simultaneous articulating faces (one articulated an /a/, the other articulated an /i/) while a naturally timed auditory vowel (either /a/ or /i/) was delivered. They found that infants looked longer at the video that matched the auditory vowel. In contrast, at that age, infants were not able to match pure tones to a corresponding video (Kuhl & Meltzoff, 1984) and also failed to do so when the tones had a distinct pitch that corresponded to a particular vowel (i.e., low tones represent auditory /a/, high tones represent /i/) or when the sounds were three-tone complexes (Kuhl, Williams, & Meltzoff, 1991, experiments 2 and 4). Based on these results, Kuhl and colleagues (1991) argued that infants' detection of AV correspondence hinges on whether or not the non-speech sound contains sufficient information to be identifiable as speech.

Others, however, have demonstrated that infants may not necessarily need to extract phonetic knowledge from visual speech in order to match the auditory and visual speech signals. For example, 7.5 month-olds can separate an auditory speech target from a distractor based on a previously seen non-speech visual signal synchronized with the auditory speech input (i.e., a squiggly horizontal line resembling an oscilloscope pattern conveying both the temporal and energetic information of the sound, Hollich, Newman, & Jusczyk, 2005). It thus seems possible that when infants hear speech, they rely on multiple cues to match it to the corresponding visual speech. According to this view, infants' detection of AV speech correspondence may be based on the correlation between the signals (e.g., Dodd, 1979; Hollich et al., 2005; Hyde, Jones, Flom, & Porter, 2011; Lewkowicz, 2000, 2010), a process that is different from matching based on phonetic information in one or both signals.

Importantly, naturalistic human speech is multisyllabic and may thus contain sufficient temporal and energetic cues that infants are able to use. If so, the Kuhl et al. findings (e.g., 1982, 1991) showing that infants cannot match single vowel-like non-speech sounds to an articulating face may not generalize to multisyllabic non-speech tokens.

Here, we sought to identify the relative weight of these cross-modal cues during infants' and adults' detection of AV speech correspondence. To do so, we used two trisyllabic AV stimuli (pseudo-words) in which the sound was either natural speech or SWS. Natural speech contains all the cues (temporal, energetic, phonetic) for matching the auditory and visual signal. In contrast, SWS contains the temporal and critical energetic information of natural speech, but phonetic detail is degraded, usually leading it to be perceived as non-speech (e.g., Remez et al., 1981).

For adults (experiment 1), we used a forced choice matching task (i.e., which of the two faces matches the audio?) and presented either natural speech sounds or SWS without making any reference to the fact that the

SWS sounds were derived from speech. For the infants (experiment 2) we used a preferential looking procedure. Although we are aware that development of the speech system undergoes significant changes in the first year of life, this study represents an initial attempt to establish the relative importance of different cross-modal perceptual cues in the AV signal for infant versus adult audiovisual speech integration. We therefore included infants across a broad age-range rather than focusing on a particular age-group.

We hypothesized that adults would perform worse with SWS than with natural speech because only natural speech contains all the phonetic details that expert listeners (i.e., adults) explicitly rely on to detect AV correspondence. For infants, we hypothesized that, if the ability to extract phonetic content from the AV signal indeed develops over time, the difference in their performance with SWS and natural speech would be smaller than in adults.

2. Experiment 1: adults

2.1. Materials and methods

2.1.1. Participants

52 undergraduate students (Mean age = 19.5 years) from the University of Connecticut participated in return for course credits after giving their written informed consent. Participants were assigned to either the natural speech (NS) or the SWS group ($N = 26$, 13 females in each group). All participants were fluent in English and in the NS group, one participant was a native speaker of Spanish and one was a native speaker of Vietnamese, while in the SWS group, two participants were native speakers of Chinese.

2.1.2. Stimuli

Stimulus creation began with a set of audiovisual recordings of a female native speaker of Dutch pronouncing two three-syllable CV-strings that made up the pseudo-words 'kalisu' and 'mufapi' (/ka/ and /fa/ as in '*ca*r' and '*fa*ther'; /li/ and /pi/ as in '*lea*n' and '*peace*'; /su/ and /mu/ as in '*soo*the' and '*moo*se'). From these recordings, we extracted two AV segments, one for 'kalisu' and one for 'mufapi', of which the audio and video signals were used to create the stimulus materials. The audiovisual phonetic contrast between the two stimuli was maximized by using opposing vowels in corresponding syllable positions and selecting consonants from different viseme-classes across stimuli (i.e., /k/ vs. /m/, /l/ vs. /f/ and /s/ vs. /p/, see Jeffers & Barley, 1971). The audio recordings were cut-off at onset and background noise was removed with the Adobe Audition 3.0 software. The overall duration of the two sounds was comparable (1028 ms for 'kalisu' and 1029 ms for 'mufapi'). Both speech signals were converted into three-tone SWS stimuli (replacing F1, F2 and F3 by sine-waves, see Fig. 1) using a script from C. Darwin (http://www.biols.susx.ac.uk/home/Chris_Darwin/Praatscripts/SWS) run in Praat, a speech analysis/synthesis software (Boersma & Weenink, 2005).

The videos showed the speaker's face from throat to crown against a dark background. They were converted into (full-color) bitmap sequences (29.97 f/s), matched on total duration (46 frames, ~1535 ms) and matched on auditory onset of the first syllable (at frame 5, see Fig. 1). During the experiment, we included six additional frames for fade-in (3 frames) and fade-out purposes.

The incongruent natural AV stimuli (e.g., hearing 'kalisu' while seeing 'mufapi' and hearing 'mufapi' while seeing 'kalisu') yielded strong perceptual mismatches due to the large contrast in phonemes and visemes. There were two inter-stimulus differences in terms of timing: the onset of the second syllable in 'kalisu' (i.e., /li/) lagged the onset of /fa/ in 'mufapi' by 16 ms whereas the onset of the third syllable in 'kalisu' (/su/) was 229 ms earlier than the onset of /pi/ in 'mufapi' (see Fig. 1).

These internal timing differences could serve as a temporal cue to the mismatch between the sound and the incorrect video. In particular, the temporal incongruence at the third syllable onset was potentially salient, given that it is larger than the temporal window of integration typically reported for adults (e.g., Grant et al., 2004; van Wassenhove et al., 2007; Vatakis & Spence, 2006).

During the experiment, a custom script was run using E-prime 1.2 software that pre-loaded the bitmap strings and allowed sounds to be delivered by trigger, ensuring natural timing without any noticeable jitter or fading. All sounds were set at an output level of 66 dBA, measured at ear-level with a Brüel & Kjær 2239 sound level meter. Two example trials can be downloaded here: http://www.martijnbaart.com/KalisuMufapiExampleTrials.zip.

2.1.3. Procedure and design

Participants were seated in a sound-attenuated and dimly lit booth in front of a 19-in. TFT screen (60 Hz refresh rate). A regular keyboard was used for data acquisition. Sounds were delivered through regular computer speakers centered beneath the screen. During an experimental trial, the two videos were presented simultaneously, one on the left side, the other on the right, while a naturally timed sound (natural speech in the NS group and SWS in the SWS group) that matched one of the two videos was delivered. The counterbalancing of sound identity ('kalisu' or 'mufapi') and side of matching video (left or right) yielded 4 different conditions. There were 48 trials in total, with 12 repetitions for each of the 4 conditions, all delivered in random order. After stimulus presentation, participants were asked to indicate whether the sound matched the left or right video by pressing a corresponding key. The next trial appeared ~1000 ms after a key-press. Importantly, the instruction made no reference to the fact that SWS sounds were derived from speech.

2.2. Results and discussion

The proportion of 'correct'-responses (i.e., when the selected video corresponded with the sound) was averaged across all 48 trials for each participant. As can be seen in Fig. 2, participants in the natural speech group quickly reached ceiling and the overall proportion of correct responses was .96, whereas participants in the SWS group performed significantly worse (mean proportion correct-

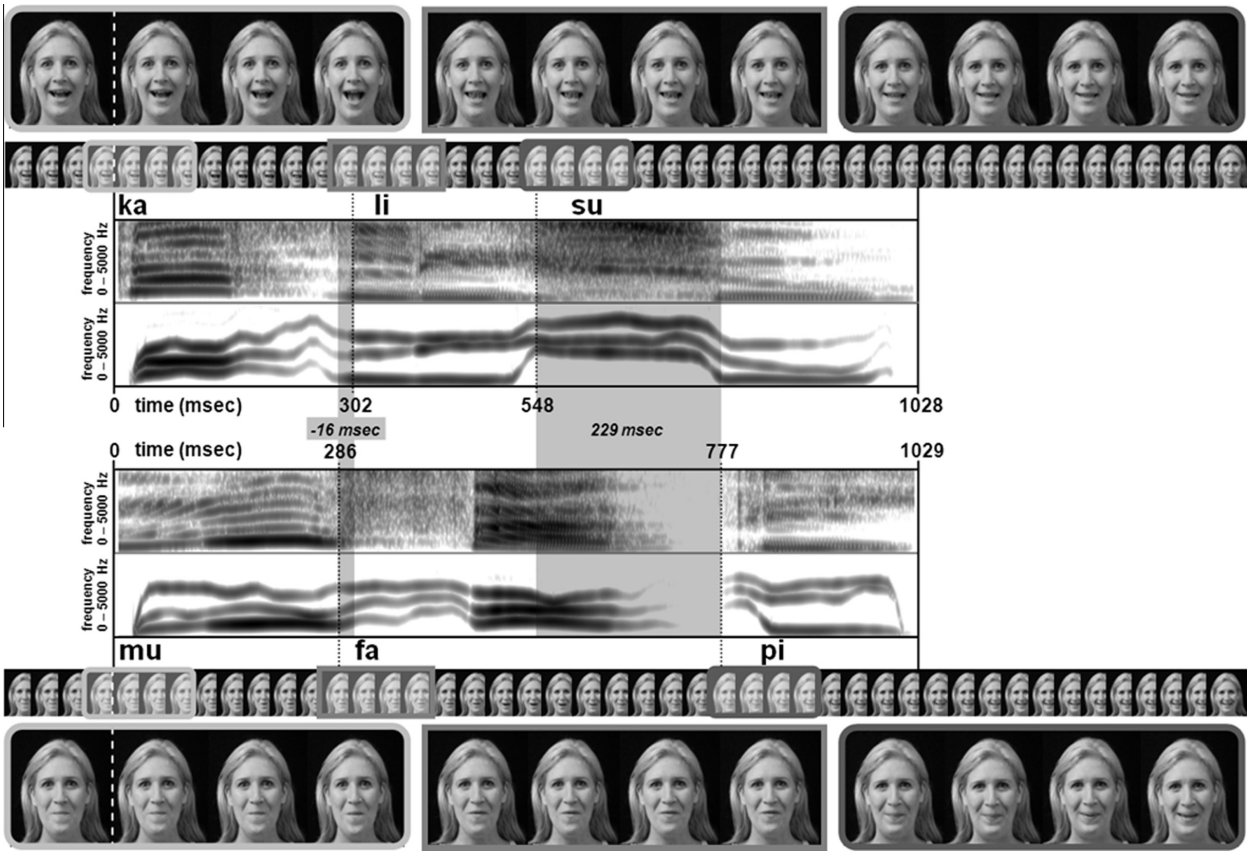


Fig. 1. Schematic overview of the audiovisual 'kalisu' (upper panels) and 'mufapi' (lower panels) stimuli. The visual speech input is depicted by individual bitmaps (i.e., 46 frames). The enlarged sections correspond to auditory onset of the three syllables in both stimuli and the white dotted line indicates auditory onset of the first syllable. The middle sections display the spectrograms of the natural speech (upper half) and SWS (lower half) sounds for both 'kalisu' and 'mufapi', relative to the timing of the visual speech. The dotted lines in the spectrograms indicate auditory onset of the second and third syllables and the grey areas indicate the timing difference in onset of the auditory syllables in 'mufapi' relative to 'kalisu'.

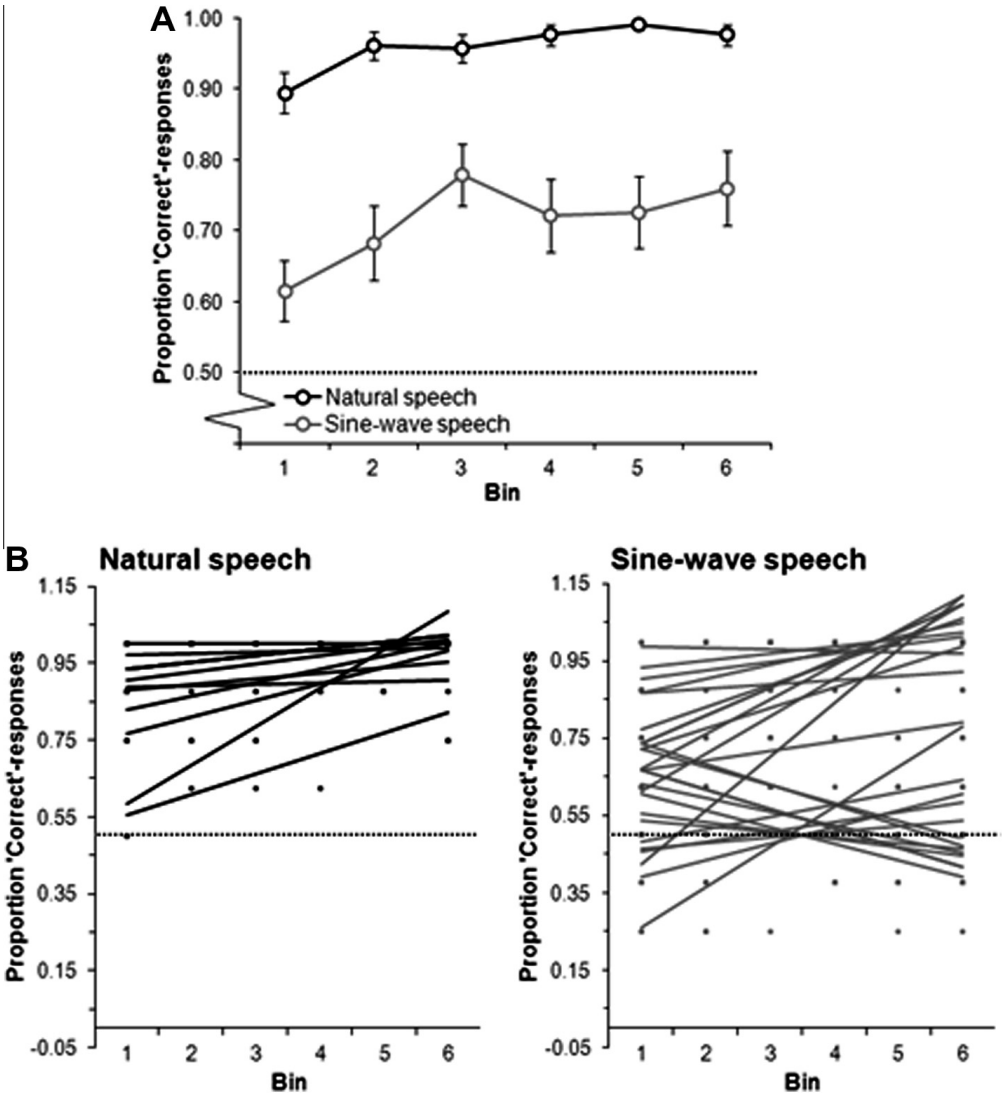


Fig. 2. Binned proportions of correct AV matches averaged across adults in the natural speech and SWS group (A) and for all adults individually (B). Error bars in 2A represent one standard error of the mean and the linear functions in Fig. 2B represent linear trend-lines fitted on the individual data. The dotted lines indicate 50% chance level. Please note that the trend-lines are included to graphically illustrate the spread and development of performance across time and are not analyzed.

responses was .71, $t[50] = 6.07$, $p < .001$) and did not reach ceiling during the test.

To examine whether there was a learning effect over the course of the experiment, we averaged the data into six bins containing 8 consecutive responses each (see Fig. 2A). A 2 (Speech type; natural speech vs. SWS) $*$ 6 (Bin) mixed-effects repeated measures ANOVA on the binned proportions of correct responses showed a main effect of Bin ($F[5,250] = 5.67$, $p < .001$, $\eta p^2 = .10$) as the proportion of correct responses increased from .75 in Bin 1 to .82 in Bin 2 ($t[51] = 2.19$, $p < .034$) and .87 in Bin 6 ($t[51] = 3.67$, $p < .001$) respectively. There was no statistical difference between Bin 6 and Bin 2, 3, 4 or 5 (p-values > .118). The ANOVA confirmed the main effect of Speech type ($F[1,50] = 36.90$, $p < .001$, $\eta p^2 = .43$) and there was no interaction between the two factors ($F[5,250] = 1.21$, $p = .304$) indicating that, on average, the speech- and SWS group improved equally over time.

It is also worth noting that the spread in performance was much larger with SWS than with natural speech. This is clearly visible in Fig. 2B, where the individual binned proportions of 'correct'-responses are depicted. Individual two-tailed binomial tests that tested the number of observed 'correct'-responses against chance-level (i.e., 24 'correct'-responses) showed that 13 participants in the SWS group did not statistically differ from chance (z-values < 1.45, p's > .05), whereas all participants in the natural speech group performed well above chance (z-values > 2.59, p's < .05). Five participants in the SWS group had given an equal number of 'correct'- and 'incorrect'-responses (they were exactly at 50% correct). Given the nature of the task, these results could potentially be caused by giving only left- or right-responses during the entire experiment. However, this was not the case as the percentage of 'left'-responses ranged between 45.8 and 62.5 for these participants.

These results thus show that adults were well able to match natural speech sounds to corresponding articulating faces and were better able to do so than adults who heard SWS (with a difference of 25% given the proportions of .96 and .71 for the natural speech and SWS groups respectively). It is likely that the detailed auditory-based phonetic content, available in natural speech only, facilitated an increase in correspondence detection.

We observed a clear distinction between SWS participants who performed above chance-level and those who performed at chance-level (none of the participants performed below 50% correct). This raises the question of whether the high performing SWS participants perceived the stimuli as more speech-like than the low performers. If so, it may reflect a top-down process that informed the listeners about the general linguistic connection between the articulating face and co-occurring auditory signal, or it could imply that whatever reduced phonetic detail that remains in SWS sounds was indeed perceived by the high-performing SWS participants (i.e., when participants partially identified one or more syllables). These possibilities were not further explored in detail because none of the SWS participants reported having identified the SWS stimuli themselves as 'kalisu' or 'mufapi' (we did not collect information regarding partial phonetic identification), which is in line with previous reports that participants need explicit training to perceive the exact identity of the SWS sounds, (e.g. Vroomen & Baart, 2009; Vroomen & Stekelenburg, 2011). The ~7% difference in performance between the high-performing SWS participants and the natural speech group (.89 versus .96 respectively, $t[37] = 2.62$ $p < .013$) might thus reflect that the natural speech group did detect, and benefitted from, the detailed AV phonetic correspondence whereas listeners in the SWS group did not, or did only partially.

We additionally observed that AV matching improved slightly over time for both groups. However, this effect is difficult to interpret since participants in the natural speech group quickly reached ceiling, leaving little room for further improvement. In contrast, participants in the SWS group did have room to improve their performance, but did not show larger learning effects than listeners in the natural speech group.

To summarize, the data suggest that AV speech integration in adults is based on multiple cues at different levels in the processing hierarchy (Eskelund et al., 2011; Schwartz et al., 2004). While adults were, in general, able to detect AV correspondence for speech-(like) material, it seems likely that most would benefit from having access to the specific phonetic content available in natural speech.

3. Experiment 2: infants

3.1. Materials and methods

3.1.1. Participants

36 infants ranging from 5 to 15 months of age participated in the experiment. Infants were randomly assigned to either the natural speech (NS) group or the SWS group ($N = 18$ in both groups). Two infants (one in the speech group and one in the SWS group) were excluded from analyses due to background noise coming from a room adjacent to the testing booth. Mean age in the final sample of 34 infants (16 females) was 9.2 months (S.D. = 2.46) and the age-distribution was alike across groups, $t[32] = .90$, $p = .373$). 28 infants were monolinguals (English) and six were raised bilingually (English/Spanish).

3.1.2. Stimuli

Stimulus material was the same as in Experiment 1.

3.1.3. Procedure and design

Infants were comfortably seated on a caregiver's lap in a dimly lit testing booth. Infants sat approximately 100 cm from two 19-in. TFT screens (60 Hz refresh rate) used for stimulus presentation, which themselves were placed 5 cm apart in a 170°-angle. Caregivers were instructed not to speak and to refrain from moving as much as possible during the experiment. The experiment was run with the E-prime 1.2 software from a laptop (Dell Latitude E4310) that controlled two video screens. The videos were 17(H) × 14(W) cm in size and spacing between the centers

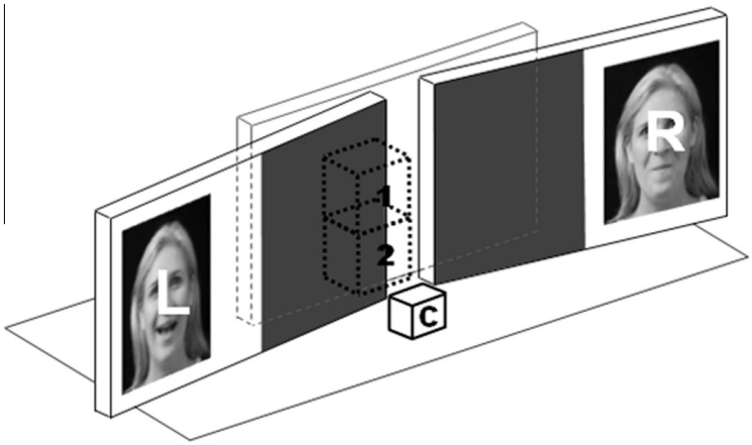


Fig. 3. Overview of the experimental set-up for infants. The left- and right screen presented the stimuli and were placed in front of the middle screen that was used to direct gaze towards midline. Looking behavior was recorded with a camera (c) and speakers 1 and 2 presented sound during fixation and stimulus presentation respectively.

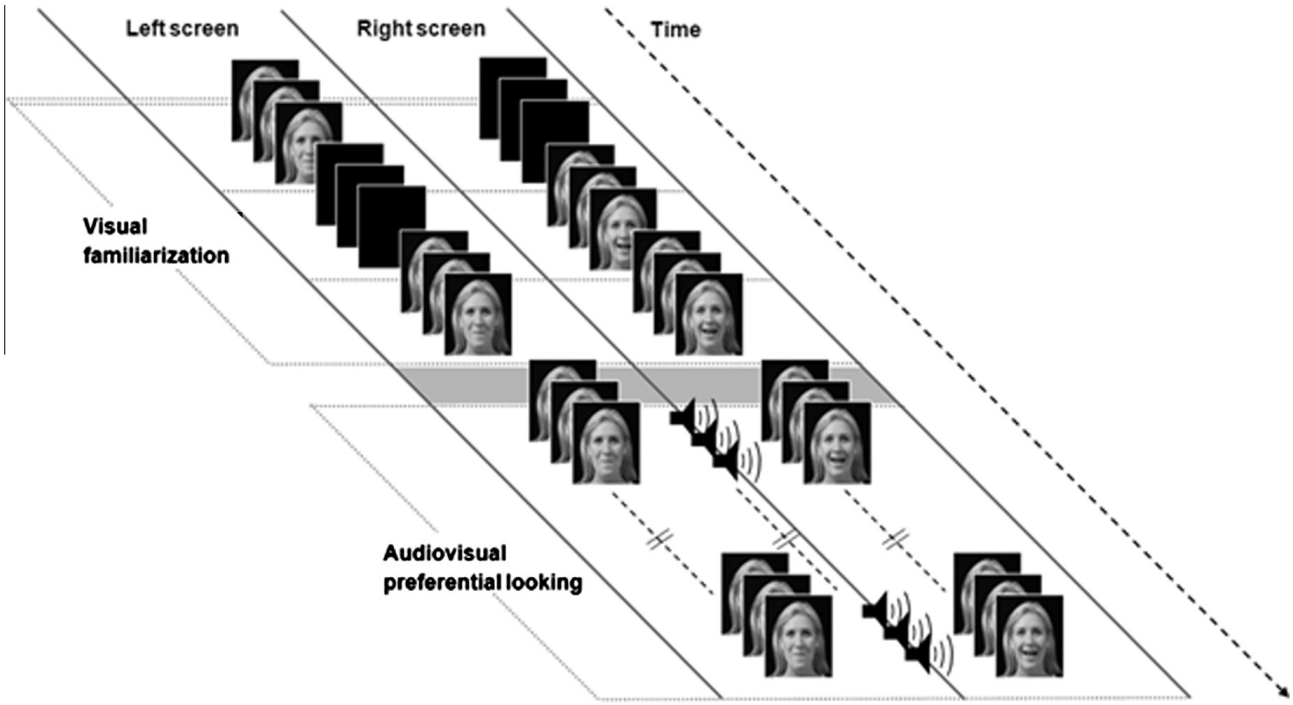


Fig. 4. Overview of a preferential looking trial. Familiarization consisted of three visual-only repetitions of a lip read video one screen, followed by three repetitions of the second video on the other screen and three simultaneous presentations on both screens. After a 1750 ms interval, both videos were displayed 36 times while a speech token that matched one of the two articulating faces was delivered.

of the left- versus right articulating mouths was 65 cm. A third TFT screen (placed behind the two screens that presented the stimuli) displayed an initial fixation stimulus and was controlled by a PC. Speech sounds were delivered through a regular PC speaker that was placed behind the screens and a second speaker delivered sounds during fixation. Infants' looking behavior was recorded by a digital video camera (Canon FS300) that was centered between the front screens (see Fig. 3).

The experiment was ~2 min in duration and consisted of three phases: a fixation phase to direct the infants' gaze towards midline, a visual-only familiarization phase to acquaint them with the display, and an audiovisual preferential looking procedure. Sound identity ('kalisu' or 'mufapi'), location of visual familiarization start (left or right screen), speech type (natural speech or SWS), and location of the matching video during testing (left or right) were counterbalanced across participants.

3.1.3.1. Fixation. A color-alternating bitmap string (~817 ms) comprised of seven images of a geometrical shape (i.e. a circle, triangle or square) was repeatedly presented. Image size was slightly increased across bitmaps to induce apparent motion towards the infant. In addition, three repetitions of an attractive sound (i.e. a squeeze-toy sound, a bicycle bell or a toy-car honk) were delivered. Fixation continued until a live feed from the camera confirmed that infants' attention was directed towards midline.

3.1.3.2. Visual-only familiarization. Infants were familiarized with the dual-screen procedure by being exposed to one video ('kalisu' or 'mufapi') on either the left or the right screen a total of three times (ISI = 500 ms), while the other screen was black (see Fig. 4). Next, the other video was displayed on the opposite side following the same procedure. Finally, three repetitions of both videos were delivered simultaneously on both screens followed by a 1750 ms period in which both screens were black.

3.1.3.3. Preferential looking. Both videos were presented simultaneously 36 times (i.e. 36 trials, ITI = 500 ms) in the same location as during familiarization, while a naturally-timed sound was played (natural speech or SWS) that matched one of the two videos. During the preferential looking procedure, looking behavior was recorded with a camera (see Fig. 3).

3.2. Results and discussion

For each infant, the camera footage obtained during the experiment was stored for off-line analysis. Frame by frame inspection and coding of all footage was done by KS, and 88% of the footage (i.e., footage from 30 infants) was coded again (by MB). Based on the frame-rate of the recordings (29.97 f/s), the number of frames that infants did/did not look at the screens were converted into milliseconds. Looking behavior during the preferential looking procedure was calculated twice: (i) from onset of the first- to offset of the last speech sound (to analyze overall looking behavior and determine overall inter-observer reliability) and (ii) from sound onset to offset for each of the 36 sound presentations (to track inter-observer reliability and infant looking behavior over time).

Inter-observer reliability was assessed by computing inter-observer Spearman's rank order correlations for the

overall time spent looking at the matching screen, the non-matching screen and time spent not looking at the screens (all ρ's > .82, p-values < .001). Additionally, we determined inter-observer reliability in more detail by calculating Cohen's Kappa for all 36 sound presentations. To do so, we categorized looking behavior during individual sound presentations as 'correct' or 'incorrect' (i.e., we categorized looking behavior as 'correct' when the time spent looking at the matching screen during a sound presentation was longer than the time spent looking at the non-matching screen). The analyses showed high inter-observer agreement during all individual sound presentations (all K's > .71, p-values < .001) and based on these findings, we averaged the double-coded data across observers for the remainder of the analyses.

Next, we determined whether the infants were actually engaged in the looking task as intended, which was indeed the case as the average proportion of time spent looking at the screens was 84% for both the natural speech group and the SWS group (t[32] = .10, p = .921).

Initial analyses showed that there was no correlation (computed separately for the speech and SWS group) between age and looking times towards the matching and non-matching screens (r-values in between −.28 and .34, p's > .185).[1] Likewise, gender, speech sound identity, and starting-location of the visual familiarization phase were not correlated with looking behavior in neither group of infants (r-values in between −.14 and .20, p's > .449). The data were therefore collapsed across these factors. There was an overall preference to look at the right- rather than the left screen, but the time spent looking at the Right vs. Left screen did not interact with the critical factor of Speech type (i.e., natural speech versus SWS), nor was there an interaction between time spent looking at the Left/Right screen, Speech type and Location of the screen (Left/Right) that matched the audio (p-values obtained in the ANOVA >.591, ηp^2 < .01). Given this, we also averaged across Location of the matching screen.

Next, we calculated the proportion of time spent looking at the screen that matched the audio (from onset of the first sound to offset of the last sound), given that infants were looking at either of the two screens (i.e., PTM for Proportion of Time spent looking at the Matching screen = time spent looking to the match/[time spent looking to the match + time spent looking to the no match]). The averaged PTM was .71 in the natural speech group versus .65 in the SWS group. As can be seen in Fig. 5, there was one infant (in the SWS group) with a PTM of 1. Although potentially, this infant may not have detected that there was visual information presented on the other (i.e., non-matching) screen, this seems highly unlikely as the infant did look at both screens during familiarization. We therefore did not exclude this infant from the analyses.[2]

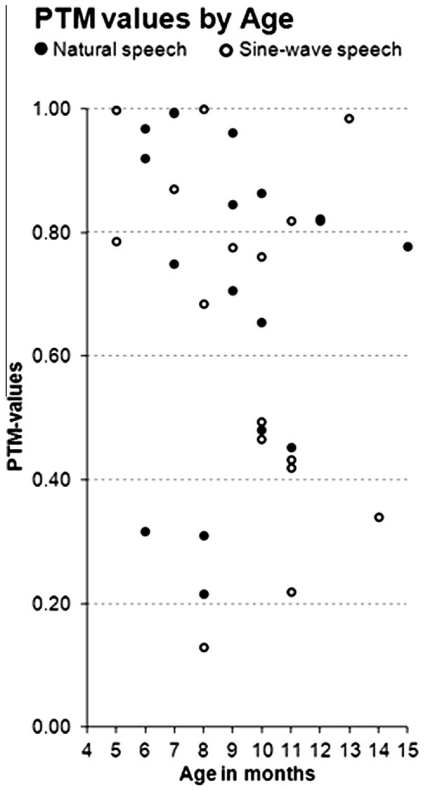


Fig. 5. PTM-values (time spent looking at the matching screen/[time spent looking at the matching screen + time spent looking at the non-matching screen]) for all infants in the natural speech (black dots) and SWS group (white dots).

Both PTM-values were higher than 50% chance level (t[16] = 3.30, p < .005, 95% CI = .57–.84, d = .801 for natural speech infants and t[16] = 2.19, p < .044, 95% CI = .50–.79, d = .532 for the SWS infants, respectively), and, most important for current study, there was no statistical difference between the natural speech- and SWS infants (t[32] = .66, p = .513, see Fig. 5 for individual data).

Next, we analyzed looking behavior over time by averaging the proportion of looking times to the matching- and non-matching screens across 6 consecutive sound presentations (from sound onset to offset for each of the six sounds), and calculating the PTM-values in each of these bins (i.e., BPTM for Binned Proportion of Time spent looking at the Matching screen). A 2 (Speech type; natural speech vs. SWS) * 6 (Bin) mixed-effects repeated measures ANOVA on the BPTM values showed a main effect of Bin (F[5, 160] = 2.30, p < .048, ηp^2 = .07) as the average BPTM increased somewhat during the first 5 Bins (30 sound presentations). As can be seen in Fig. 6A, the average BPTM across conditions was highest in Bin 5 (.73) and lowest in Bin 1 (.60, t[33] = 2.46, p < .020), indicating an overall learning effect that marginally dropped off during the last 6 sound presentations (i.e., the difference between BPTM-values in Bin 5 vs. Bin 6 was hovering on significance, p = .051).

[1] This was corroborated by two ANOVAs on the time spent looking at the matching/non-matching screens including the factors Speech Type (natural speech or SWS) and Age (Age was included either as a covariate or as a 2-level between-subjects factor through a median split) that yielded p's > .260 for the main effects of Age and interactions involving Age.

[2] All PTM (and subsequent BPTM) analyses yielded the same statistical results when the infant with a PTM of 1 was excluded.

Most importantly, the ANOVA showed no effect of Speech type ($F < 1$) and no interaction between Speech type and Bin ($F[5,160] = 1.41$, $p = .223$; additional independent-samples t-tests that tested BPTM between groups in each bin all yielded p's > .345), indicating that looking behavior developed equally over time across the two groups of infants.

As indicated in Fig. 6B, the spread in performance was alike for both groups of infants. To gain more insight into the spread, we conducted two-tailed binomial tests in which we tested the observed amount of 'correct' sound presentations (during which infants looked longer at the matching- than non-matching screen) against chance-level (i.e., 18 sound presentations with 'correct' looking behavior). The analyses showed that 3 infants in the speech group and 3 infants in the SWS group performed below chance (the observed amount of 'correct' sound presentations $\leqslant 12$, z's $\geqslant 2$, p's < .05). There were 5 infants in the natural speech group and 4 infants in the SWS group that did not differ from chance-level (observed 'correct' amount of sound presentations $\geqslant 13$ and $\leqslant 23$, z's < 1.67, p's > .05) and the majority of the infants (9 in the natural speech group and 10 in the SWS group) performed above chance-level (observed 'correct' amount of sound presentations $\geqslant 24$ z's > 2, p's < .05).

The data thus show that infants preferred to look at the screen that matched the auditory input, irrespective of whether the sounds were natural speech or artificial SWS. We observed no group differences on the (binned) PTM-values and looking behavior in the two groups of infants developed equally during the course of testing. Overall, we observed two contrasts between adults and infants, namely, (i) adults performed worse on AV matching with SWS as compared to natural speech whereas infants performed equally well with both speech types and (ii) the spread in the infant data was similar for both speech types whereas the SWS adults showed larger variability than natural speech adults.

4. General discussion

The current study showed that adults asked to match natural speech with corresponding visual speech performed better than adults asked to match the SWS

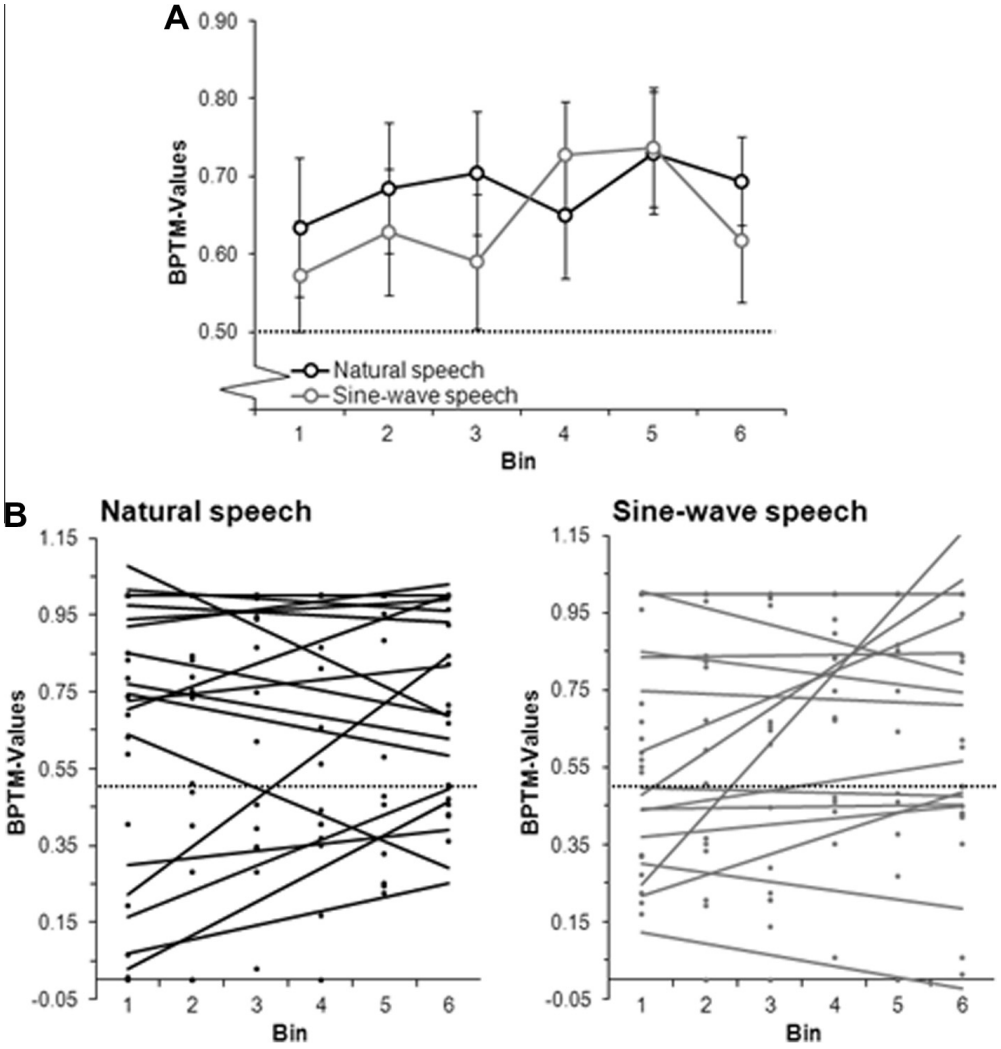


Fig. 6. Binned proportions of looking at the matching screen (BPTM) averaged across infants in the natural speech and SWS group (A) and for all infants individually (B). Error bars in 6A represent one standard error of the mean and the linear functions in Fig. 6B represent linear trend-lines fitted on the individual data. The dotted lines indicate 50% chance level. Please note that the trend-lines are included to graphically illustrate the spread and development of performance across time and are not analyzed.

counterparts to the same articulating faces, whereas infants matched the auditory signal with the correct articulating face irrespective of whether the sounds were speech or SWS.

We proposed that observers may use some combination of temporal, energetic, and phonetic cues to match AV speech. In our stimuli, the most prominent temporal cue was the AV asynchrony between the sound and the non-matching visual speech. It is well documented that both adults and infants are sensitive to AV (a)synchrony in both speech- and non-speech stimuli (e.g., Bahrick, 1983, 1987; Dodd, 1979; Grant et al., 2004; Hyde et al., 2011; Lewkowicz, 1986, 1996, 2000, 2003, 2010; Pons, Teixidó, Garcia-Morera, & Navarra, 2012; van Wassenhove et al., 2007; Vatakis & Spence, 2006; Vroomen & Stekelenburg, 2010, 2011) but since adult perception of AV synchrony is alike for natural speech and SWS (Vroomen & Stekelenburg, 2010, 2011, see Grant et al., 2004 for similar findings with artificial speech comprised of spectrally distinct 1/3-octave bands), AV asynchrony was a consistent cue for all adults.

For infants who heard 'kalisu', the auditory onset of /su/ led the corresponding segment of the 'mufapi' video by less than 229 ms; for infants who heard 'mufapi', the auditory onset of /pi/ lagged the corresponding segment of the 'kalisu' video by more than 229 ms. Nonetheless, we observed no correlation between speech sound identity and the PTM-values. It therefore seems likely that the cue of AV asynchrony was equally important to all infants. Moreover, a recent study (Lewkowicz, 2010) identified a 300 ms asynchrony threshold for AV speech in 4-, 6-, 8-, and 10-month-old infants. That is, when the infants were familiarized to a 666 ms asynchrony (in which the audio led the visual signal) they could detect the change from this asynchrony to a 366 ms auditory lead asynchrony, but not to a 500 ms asynchrony (i.e., a 166 ms difference between familiarization and test). However, whenever the familiarization comprised synchronous AV presentations, infants did not detect the 366 ms asynchrony during test. Based on these findings, it seems that the infants in our study may indeed not have perceived the asynchrony between the sound and the mismatching video. Interestingly, Lewkowicz (2010, page 73) noted that infants typically are not exposed to asynchronous AV speech and thus may be biased towards synchronous AV events. According to this view, the perfect temporal correlation between the sounds and their matching videos may have provided infants with a sufficiently strong cue, consistent with the default they experience in all natural AV speech events.

Moreover, it is known that infants' auditory system is sensitive to 25 ms temporal modulations in concatenated noise segments (Telkemeyer et al., 2009), which corresponds to the temporal modulations needed to extract segmental information from the speech signal (Rosen, 1992). Specifically, when comparing 12 ms modulations with 25 ms modulations, Telkemeyer and colleagues (2009) reported enhanced brain activity for the latter specifically in neonates' bilateral inferior and posterior temporal brain regions, as well as the in the right temporoparietal region, a brain area demonstrated to be sensitive to auditory sequences with temporal structure similar to speech syllables (Homae, Watanabe, Nakano, & Taga, 2012). We therefore assume that all infants were able to detect the temporal auditory structure in the stimuli in detail, and we propose that they could use the cross-modal temporal correlation to match the sound with the video. This inference seems plausible when considering that infants cannot match a 1.12 s or 1.16 s static artificial three-tone complex onto visual speech (Kuhl et al., 1991), whereas, as demonstrated here, infants can match three-tone complexes that share the temporal relationship that exists between natural speech with visual speech. Our data thus indicate that both adults and infants are able to detect AV correspondence based on the *AV temporal correlation* in the signal.

Although the *correlation between acoustic energy and the visible articulators* is, by default, closely related to phonetic information (e.g., Fant, 1960; Lieberman & Blumstein, 1988), most of the acoustic energy from speech is available in its SWS analog, whereas the phonetic detail is severely compromised. To be more precise, high frequency modulations or acoustic bursts in natural /s/ and /p/ are partially filtered out in the SWS stimuli but the correlation between the area of lip opening and acoustic energy in the F2 and F3 frequency-bands (which is preserved in SWS) is demonstrated to be critical (Grant & Seitz, 2000). In contrast, none of the adults could phonetically indentify the entire SWS stimuli, indicating that the phonetic information was compromised relative to that available in natural speech.

Of particular relevance here are previous studies demonstrating that infants can use phonetic information in visual speech (e.g., Bristow et al., 2008; Burnham & Dodd, 2004; Kuhl & Meltzoff, 1982, 1984; Kushnerenko et al., 2008; Patterson & Werker, 2003; Rosenblum et al., 1997; Teinonen, Aslin, Alku, & Csibra, 2008), although not as mandatorily as adults (Desjardins & Werker, 2004).

For instance, Rosenblum and colleagues (1997) habituated 5-month-old infants to an AV speech stimulus comprised of an auditory and visual /va/ (AV_{va}) and tested them on incongruent stimuli in which the visual speech signal was /va/ (V_{va}) whereas the auditory signal was either /ba/ ($A_{ba}V_{va}$) or /da/ ($A_{da}V_{va}$). The authors showed that infants did not detect the difference between AV_{va} and $A_{ba}V_{va}$ stimuli, but did detect the difference between AV_{va} and $A_{da}V_{va}$. There are two explanations that could account for this finding. One is that infants may have perceived the visual energetic cues in the bilabial V_{va} as compatible with the bilabial and labio-dental A_{va} and A_{ba} but not with the alveolar A_{da}. The other is that infants effectively perceived the $A_{ba}V_{va}$ stimulus as /va/ because of a visually induced bias on phonetic sound identity. As was previously mentioned, the SWS stimuli employed in the current study allowed us to partially separate energetic cues from phonetic cues. Nonetheless, it seems most appropriate to consider energetic cues as an integral part of the phonetic cues in speech. As a result, we cannot conclude that observers (irrespective of whether they were infants or adults) did not use any phonetic information in the SWS to match the sound with the corresponding video, but they used substantially less as SWS simply does not contain the phonetic detail available in natural speech.

As noted, the ability to extract *phonetic information* from visual speech appears to increase over developmental

time. For instance, the visual bias on sound identification in children is reported to be less than 10%, up to a maximum of about 57%, as compared to the visual bias in adults (Desjardins et al., 1997; Hockley & Polka, 1994; Massaro, 1984; Massaro, Thompson, Barron, & Laren, 1986). The framework as laid out in the Fuzzy Logical Model of Perception (FLMP, Massaro, 1987) suggests that the unimodal perceptual inputs are compared with stored language specific representations at the final stage of AV integration (Massaro, 1998; Massaro, Cohen, Campbell, & Rodriguez, 2001). According to this view, the increasing influence of visual speech during development can be explained by a strengthening of phonetic (and other) representations that are the result of experience with language. It therefore may come as no surprise that we indeed observed that availability of the full phonetic detail led to higher performance in adults. Specifically, we obtained additional evidence in support of the multi-stage model of AV integration whereby different cues are integrated on different levels (e.g., Eskelund et al., 2011; Schwartz et al., 2004; Vroomen & Stekelenburg, 2011). Furthermore, this process appears to be hierarchical in the sense that lower-level temporal and spatial AV features are integrated *before* higher-order phonetic features (e.g., Klucharev, Möttönen, & Sams, 2003; Stekelenburg & Vroomen, 2007, 2012; van Wassenhove, Grant, & Poeppel, 2005).

For infants, however, the data-pattern may be explained in a couple of different ways. First, given that we observed no difference between the SWS and natural speech infants, one could argue that infants were able to detect AV phonetic correspondence equally well for natural speech and SWS. This would be a potentially interesting interpretation because it implies that infants can detect significantly more phonetic detail in SWS than adults. Although it is understood that infants may perceive more phonetic detail than adults (i.e., because the perceptual system narrows down towards the native language during development, e.g., Pons, Lewkowicz, Soto-Faraco, & Sebastián-Gallés, 2009; Werker & Tees, 1984), it seems unlikely that they would perceive SWS and natural speech as equally phonetically informative. Infants and neonates prefer to listen to real speech rather than to SWS (Vouloumanos & Werker, 2004, 2007) and, although it has been suggested that this may be so because the rich melodic voice pitch contour of speech is absent in SWS (Rosen & Iverson, 2007), Vouloumanos and Werker (2007) argued in favor of a biological preference for speech- over non-speech sounds. This suggestion is indirectly supported by recent data showing stronger brain activity in the left posterior temporal area for speech than for SWS (Homae, Watanabe, & Taga, 2013), a finding that aligns with earlier reports in which speech was contrasted with various non-speech sounds, silence, and backward speech, and showed that speech is processed predominantly in the left hemisphere (Bortfeld, Fava, & Boas, 2009; Bortfeld, Wruck, & Boas, 2007; Kotilahti et al., 2010; Minagawa-Kawai et al., 2011; Peña et al., 2003).

In addition, Desjardins et al. (1997) demonstrated that the perceptual influence of visual speech may depend on how well children are able to produce speech themselves (i.e., children who made substitution errors while producing consonants were less influenced by the visual speech signal than children who did not make these production errors). This is in line with work showing that the influence of visual speech on auditory perception is stronger for children with delayed speech than for children with truly disordered phonology (Dodd, McIntosh, Erdener, & Burnham, 2008). Finally, it has been demonstrated that infants' internal memory representations of speech sounds lead to a perceptual clustering around prototypes of these sounds (e.g., Kuhl, 1991; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992), which then may serve as targets for speech production (Kuhl & Meltzoff, 1996). All of these findings are relevant to the current study as they indicate that only the perception of natural speech, and not SWS, is linked to subsequent speech production.

The second possibility is that the relative contribution of phonetic and non-phonetic cues used by infants is variable and depends on the relative saliency of AV stimulus properties. To explain this in detail, we again need to start from the observation that we found no statistical difference between infants who heard SWS and those who heard natural speech. This is in contrast to Kuhl and Meltzoff's (1982) observation that infants' (18 to 20 weeks old) proportion of looking times to the face that articulated an auditory vowel dropped from ~74% to chance (~55%) when pure-tone stimuli were used. This could simply be because SWS contains more phonetic information than pure-tones. However, Kuhl and colleagues (1991) subsequently observed that infants could not match three-tone vowel analogues—quite similar to SWS—with corresponding visual speech. Apparently, the SWS-like nature of a sound as such (i.e., when critical formant center frequencies are retained) does not necessarily imply that AV phonetic correspondence detection will occur. Here, we used CVCVCV SWS stimuli, from which infants could extract more cues than they were able to from artificial three-tone vowels.

Regardless, the fact that both groups of infants tested here performed alike indicates that they do not need robust phonetic information to detect AV correspondence in speech-like material. Of course, this does not imply that infants cannot, and do not, use phonetic cues. In fact, although not significant, the average PTM was somewhat higher (~6%) when the sounds were natural speech rather than SWS. This hints at the possibility that phonetic information has *a small additional benefit when it is redundant*, (i.e., when there are salient lower level non-phonetic cues, as presumably is the case here) whereas it has a *large correspondence detection benefit when it is the most prominent available cue* (i.e., in single vowels where non-phonetic cues are virtually absent e.g., Kuhl and Meltzoff, 1982; Kuhl et al., 1991).

Follow-up studies are needed to examine this suggestion in more detail. In particular, the use of more sensitive paradigms or procedures that move beyond solely behavior-based indications of perceptual matching (e.g., EEG, NIRS) may provide more nuanced evidence of how infants process of the two stimulus types. Based on the current findings, it seems appropriate to conclude that (i) linguistic experience increases the perceptual weight of phonetic cues in the AV signal and (ii) infants' use of the available AV cues is more variable than adults'. Future research will provide

additional insights by (i) systematically degrading the speech signal in time, energy and phonetic detail, (ii) focusing on within-infancy development by including infants from specific age groups, and (iii) incorporating more sensitive measures to chart the developmental trend underlying the relative contributions of these cues for speech perception.

5. Conclusion

Our data indicate that adults' detection of AV speech correspondence improves when the availability of phonetic detail in the signal is increased. In contrast, infants' performance was not affected by increases in the amount of phonetic information.

The data corroborates accounts of a multi-stage AV integration process in adults, and suggests that infants' matching of audio and visual speech can be driven by salient non-phonetic properties of the signal. This is particularly important, given different accounts of when phonetic audiovisual integration emerges during development. Findings that infants match audio speech to visual speech based on phonetics (e.g., Kuhl and Meltzoff, 1982; Patterson and Werker, 1999, 2003) conflict with those demonstrating that the ability to extract phonetic information from AV speech increases with development (e.g., Massaro, 1984; McGurk and MacDonald, 1976). Our data suggest that infants use AV cues to detect correspondence between a sound and visual speech more flexibly than adults. This supports an account by which the relative weights of different perceptual cross-modal cues change across development, such that adults rely more on the phonetic content of the stimuli than infants.

Acknowledgements

This research was supported by National Institutes of Health Grant R01 DC010075 to Heather Bortfeld, a Fulbright grant for PhD.-students to Martijn Baart and a National Science Foundation IGERT Grant 1144399 to the University of Connecticut.

References

Bahrick, L. E. (1983). Infants' perception of substance and temporal synchrony in multimodal events. *Infant Behavior and Development, 6*(4), 429–451.

Bahrick, L. E. (1987). Infants' intermodal perception of two levels of temporal structure in natural events. *Infant Behavior and Development, 10*, 387–416.

Boersma, P., & Weenink, K. (2005). Praat: Doing phonetics by computer. <http://www.fon.hum.uva.nl/praat>.

Bortfeld, H., Fava, E., & Boas, D. A. (2009). Identifying cortical lateralization of speech processing in infants using near-infrared spectroscopy. *Developmental Neuropsychology, 34*(1), 52–65.

Bortfeld, H., Wruck, E., & Boas, D. A. (2007). Assessing infants' cortical response to speech using near-infrared spectroscopy. *Neuroimage, 34*(1), 407–415.

Bristow, D., Dehaene-Lamberts, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., et al. (2008). Hearing faces: How the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience, 21*(5), 905–921.

Bruce, V., Campbell, R. N., Doherty-Sneddon, G., Import, A., Langton, S., McAuley, S., et al. (2000). Testing face processing skills in children. *British Journal of Development Psychology, 18*, 319–333.

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology, 45*(4), 204–220.

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.

Burnham, D., & Sekiyama, K. (2004). When auditory-visual speech perception develops: The locus of the Japanese McGurk effect. *Australian Journal of Psychology, 56*, 108.

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.

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology, 11*(4), 478–484.

Dodd, B., McIntosh, B., Erdener, D., & Burnham, D. (2008). Perception of the auditory-visual illusion in speech perception by children with phonological disorders. *Clinical Linguistics & Phonetics, 22*(1), 69–82.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*, 303–306.

Eskelund, K., Tuomainen, J., & Andersen, T. S. (2011). Multistage audiovisual integration of speech: Dissociating identification and detection. *Experimental Brain Research, 208*(3), 447–457.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America, 109*(5), 2272–2275.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108*(3), 1197–1208.

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Communication, 44*(1–4), 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*(5), 3309.

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development, 76*(3), 598–613.

Homae, F., Watanabe, H., Nakano, T., & Taga, G. (2012). Functional development in the infant brain for auditory pitch processing. *Human Brain Mapping, 33*(3), 596–608.

Homae, F., Watanabe, H., & Taga, G. (2013). Speech and sine wave speech processing in the infant brain, Poster presented at the Workshop on Infant Language Development (WILD). June 20–22, Donostia – San Sebastián, Spain.

Hyde, D. C., Jones, B. L., Flom, R., & Porter, C. L. (2011). Neural signatures of face-voice synchrony in 5-month-old human infants. *Developmental Psychobiology, 53*(4), 359–370.

Jeffers, J., & Barley, M. (1971). *Speechreading (lipreading)*. Springfield, Illinois: Charles C. Thomas.

Jusczyk, P. W., & Luce, P. A. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*(5), 630–645.

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research, 18*(1), 65–75.

Kotilahti, K., Nissilä, I., Näsi, T., Lipiäinen, L., Noponen, T., Meriläinen, P., et al. (2010). Hemodynamic responses to speech and music in newborn infants. *Human Brain Mapping, 31*(4), 595–603.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*(2), 93–107.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138–1141.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development, 7*(3), 361–381.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America, 100*(4), 2425–2438.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*(2), F13–F21.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606–608.

Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 17*(3), 829–840.

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*(32), 11442–11445.

Lewkowicz, D. J. (1986). Developmental changes in infants' bisensory response to synchronous durations. *Infant Behavior and Development, 9*, 335–353.

Lewkowicz, D. J. (1996). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology; Human Perception and Performance, 25*(5), 1094–1106.

Lewkowicz, D. J. (2000). Infants' perception of the audible, visible, and bimodal attributes of multimodal syllables. *Child Development, 71*(5), 1241–1257.

Lewkowicz, D. J. (2003). Learning and discrimination of audiovisual events in human infants: the hierarchical relation between intersensory temporal synchrony and rhythmic pattern cues. *Developmental Psychology, 39*(5), 795–804.

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology, 46*(1), 66–77.

Lieberman, P., & Blumstein, S. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge: Cambridge University Press.

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

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge: The MIT Press.

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review, 8*, 1–17.

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology, 41*, 93–113.

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

Minagawa-Kawai, Y., van der Lely, H., Ramus, F., Sato, Y., Mazuka, R., & Dupoux, E. (2011). Optical brain imaging reveals general auditory and language-specific processing in early infant development. *Cerebral Cortex, 21*, 254–261.

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology, 22*(1), 46–49.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development, 22*, 237–247.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191–196.

Peña, M., Maki, A., Kovacic, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., et al. (2003). Sounds and silence: An optical topography study of language recognition at birth. *Proceedings of the National academy of Sciences of the United States of America, 100*, 11702–11705.

Pons, F., Teixidó, M., Garcia-Morera, J., & Navarra, J. (2012). Short-term experience increases infants' sensitivity to audiovisual asynchrony. *Infant Behavior and Development, 35*, 815–818.

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America, 106*(26), 10598–10602.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society B: Biological Sciences, 336*(1278), 367–373.

Rosen, S., & Iverson, P. (2007). Constructing adequate non-speech analogues: what is special about speech anyway? *Developmental Science, 10*(2), 165–168.

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

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*, 347–357.

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*(12), 2329–2337.

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition, 93*(2), B69–78.

Sekiyama, K., & Burnham, D. (2004). Issues in the development of auditory-visual speech perception: Adults, infants, and children. *Paper presented at the 8th International Conference on Spoken Language Processing*, Jeiu Island, Korea.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*(12), 1964–1973.

Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Frontiers in Integrative Neuroscience, 6*.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Swingley, D., Pinto, J. P., & Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition, 71*, 73–108.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 108*(3), 850–855.

Telkemeyer, S., Rossi, S., Koch, S. P., Nierhaus, T., Steinbrink, J., Poeppel, D., et al. (2009). Sensitivity of newborn auditory cortex to the temporal structure of sounds. *The Journal of Neuroscience, 29*(47), 14726–14733.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96*(1), B13–22.

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*(4), 1181–1186.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598–607.

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research, 1111*(1), 134–142.

Vouloumanos, A., & Werker, J. F. (2004). Tuned to the signal: the privileged status of speech for young infants. *Developmental Science, 7*(3), 270–276.

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: evidence for a bias for speech in neonates. *Developmental Science, 10*(2), 159–171.

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition, 110*(2), 254–259.

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience, 22*, 1583–1596.

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition, 118*(1), 75–83.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49–63.