Development and Aging

Perception of audiovisual infant directed speech

NUNNE ENGLUND and DAWN M. BEHNE

Department of Psychology, NTNU, Norwegian University of Science and Technology, Trondheim, Norway

Englund, N. & Behne, D. M. (2019). Perception of audiovisual infant directed speech. *Scandinavian Journal of Psychology.*

Infant perception often deals with audiovisual speech input and a first step in processing this input is to perceive both visual and auditory information. The speech directed to infants has special characteristics and may enhance visual aspects of speech. The current study was designed to explore the impact of visual enhancement in infant-directed speech (IDS) on audiovisual mismatch detection in a naturalistic setting. Twenty infants participated in an experiment with a visual fixation task conducted in participants' homes. Stimuli consisted of IDS and adult-directed speech (ADS) syllables with a plosive and the vowel /a:/, /i:/ or /u:/. These were either audiovisually congruent or incongruent. Infants looked longer at incongruent than congruent syllables and longer at IDS than ADS syllables, indicating that IDS and incongruent stimuli contain cues that can make audiovisual perception challenging and thereby attract infants' gaze.

Key words: Perception, speech, infant.

Nunne Englund, Department of Psychology, NTNU, Norwegian University of Science and Technology, N-7491 Trondheim, Norway. Tel: +47 73 59 05 69; Fax +47 92 44 20 89. e-mail: nunne.englund@ntnu.no

INTRODUCTION

Infant-directed speech (IDS), when compared to adult-directed speech (ADS), has characteristics which makes it well suited for speech and language learning (Cristia, 2013; Weisleder & Fernald, 2013; Werker & McLeod, 1989; Zhang, Koerner, Miller *et al*., 2011). Findings indicate that as well as enhancing auditory speech cues, speakers of IDS may also enhance visual speech cues (Benders, 2013; Englund & Behne, 2005; Martin, Schatz, Versteegh *et al*., 2015). When presented with speech information to visual and auditory modalities, infants perceive whether they match or not (Kuhl & Meltzoff, 1984; Patterson & Werker, 1999; Walton & Bower, 1993). The current study explores the impact of visual enhancement in IDS on audiovisual mismatch detection. The current research may contribute to knowledge on phonological development in infants, as well as a foundation for advice on how to optimally adapt speech to infants, both with normal and abnormal speech and language development.

Audiovisual speech perception in infants

An early language environment presents considerable complexity to an infant sorting out phonetic categories, not only by its variability in speakers and speech registers, but also by the availability of information from multiple sensory modalities. Despite this complexity, new-borns are able to perceive commonalities in visual and auditory information in continuous speech (Guellaï, Streri, Chopin, Rider & Kitamura, 2016). While still not a robust effect (Desjardins & Werker, 2004), classical studies show that infants will look longer at the face articulating a heard vowel compared to a non-matching vowel (Aldridge, Braga, Walton & Bower, 1999; Kuhl & Meltzoff, 1984; Patterson & Werker, 1999). In these classical studies, infants are presented with side-by-side displays of two faces, each with an adult articulating a vowel (e.g., /i/ and/a/) while hearing only one of them. When this method is used, infants will typically look at the screen with visual information matching the sound. Although sensitivity to matching vowels has been observed in infants as young as 2 months (Patterson & Werker, 2003), the ability to associate auditory to visual cues does not mean the two are integrated (Shaw & Bortfeld, 2015). Nonetheless, one can observe signs of audiovisual integration in 4.5-month-olds (Burnham & Dodd, 2004). Integration refers to tying auditory and visual information as a joint percept and is based on audiovisual binding which refers to connecting sensory input from the auditory and visual modalities (Nahorna, Berthommier & Schwartz, 2012). While infants are able to perceive both auditory and visual information in speech, are both kinds of information beneficial to speech processing? One study indicates that visual articulation enhances auditory phoneme discrimination (Teinonen, Aslin, Alku & Csibra, 2008). Two groups of 6-month-olds were familiarised with syllables from the middle range of the auditory /ba-da/ continuum. One group of infants (the two-category group) were tested with a visual /ba/ or /da/, chosen according to whether the auditory token was on the /ba/ or /da/ side of the midpoint of the continuum. The other group of infants (the one-category group), were tested with stimuli in which every auditory token was always paired with the same syllable, either a visual /ba/ or a visual /da/. Results revealed that infants in the two-category group, but not in the one-category group, discriminated the /ba/-/da/ contrast. This demonstrates that the visual information necessary for audiovisual binding contributes to phonetic learning at 6 months old (Teinonen *et al*., 2008). Research with adults shows that the clarity of the visual component leads to better audiovisual integration (Tiippana, 2014), and if this is also the case with infants, a language environment which enhances visual speech cues will benefit audiovisual perception and consequently phonological learning.

IDS and visual cues

From starting out with the ability to discriminate all speech categories, infant speech perception is narrowed down to the speech sounds in the ambient language (Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola & Nelson, 2008). Much of the speech infants encounter during their first months is IDS, and several studies demonstrate that formant frequencies in IDS vowels are adjusted leading to an expanded vowel space (Bernstein Ratner & Luberoff, 1984; Burnham, Kitamura & Vollmer-Conna, 2002; Kuhl, Andruski, Chistovich *et al*., 1997). However, an increasing body of research challenges these findings (Benders, 2013; Englund & Behne, 2005; Martin *et al*., 2015). Englund and Behne (2005, 2006) have demonstrated that instead of a stretched vowel space, vowels in IDS were fronted; F_1 was higher in IDS than ADS for all vowel qualities tested, while F_2 was higher in IDS than ADS for /a:-a/ and/u:-u/, but not for /i:-i/. Benders (2013) found similar results in a study of IDS and ADS. She observed F_2 to be higher in IDS than ADS, for a:/ and /u:/ but not for/i:/. One of the interpretations from these studies is that IDS offers visual enhancement of speech sounds. Supporting evidence comes from a study by Green, Nip, Wilson, Mefferd and Yunusova (2010) demonstrating lip movements to be exaggerated during IDS compared to ADS. Mouth opening and vertical lip aperture were wider during IDS compared to ADS (Green *et al*., 2010). Findings were verified in analyses of formant frequencies, revealing F_1 to be higher during IDS than ADS for the vowels studied, and F_2 to be higher in IDS than ADS for the low vowels /ae/ and /a/, but not for /i/ (Green *et al*., 2010). An alternative account for different formant frequencies in IDS has recently been presented, based on eight mothers' vocal tract length estimated by formant frequency values in IDS speech (Kalashnikova, Carignan & Burnham, 2017). The authors claim that IDS vowel acoustics is a product of laryngeal raising, resulting from social convergence. Specifically, reducing the length of the vocal tract will mimic vocal tract length of an infant. This can explain generally higher formant frequencies in IDS, but cannot account for the smaller vowel space seen in studies of Norwegian IDS (Englund & Behne, 2006). Together, these findings indicate that vowels are articulated more front in the vocal tract during IDS compared to ADS, a possible reason for it being enhancement of visual speech cues. In turn, this raises the question of whether audiovisual IDS and ADS are perceived differently.

Infants will look longer at the mouth of someone speaking IDS compared to ADS (Lewkowicz & Hansen-Tift, 2012). Furthermore, discrimination studies show that 4-, 6-, and 8-month-olds are more sensitive to changes in audiovisual speech cues in IDS compared to ADS (Lewkowicz, 1996, 2000). Kubicek, Gervain, de Boisferon, Pascalis, Lœvenbruck & Schwarzer (2014) have studied the influence of IDS on infants' intersensory perception of fluent speech, showing that IDS facilitates infants' ability to match auditory to visual information in their native language. These findings imply that audiovisual IDS make perception of speech easier for an infant. Yet, results could be biased by a number of factors. One is facial cues, which are different in IDS than ADS. Speakers widen eyes, raise eyebrows and smile more during IDS, which may attract infants' attention and lead to better matching (Chong, Werker, Russell & Carroll, 2003). Another is movement. One study of 8-month-olds used point-line displays of the face or head movements of a woman producing IDS. Infants fixated on a corresponding auditory match from the kinematic information about the head but not from the same type of information about lip movements (Kitamura, Guellai & Kim, 2014). Although this study used long stretches of speech, it is noteworthy that head movement can affect audiovisual perception. Together, these studies reveal that audiovisual IDS may enhance matching of auditory and visual information. Still, more studies are needed using IDS stimuli with a smaller vowel space relative to ADS stimuli.

Experimental paradigm

Infant speech perception is commonly studied in a laboratory providing the possibility for sound proofing and an otherwise controlled environment. However, a lab can be an artificial setting for infants and parents, and parents may behave differently in a lab than at home (Stevenson, Leavitt, Roach, Chapman & Miller, 1986) which in turn can affect infants. Given the greater familiarity, a home setting may elicit more natural behaviour from participants increasing ecological validity.

Infant studies stand the risk of attrition, resulting in low external validity. Several factors can decrease this risk (Pomerleau, Malcuit, Chamberland, Laurendeau & *et al*., 1992) and the less demanding the experimental task requiring sustained attention, the more likely infants will finish the procedure (Oates, 1998). Hence, a relatively simple task may lead to low attrition. Simplicity can be expressed both in the duration of the procedure as well as in the requirements for response. A visual fixation paradigm includes a practical and simple task that can be used with infants across a wide age range, from 2- to 14-month-olds (Jusczyk, 1997). Studies of infant perception in natural settings with simple procedures are scarce and more research is needed where one explores how home environments can be used for experimental studies.

METHOD

To our knowledge, no studies have tested young infants' perception of audiovisual IDS with acoustic cues similar to those in Norwegian IDS, in particular in a natural setting. We ask whether the potential visual enhancement in audiovisual IDS is prominent for an infant to the extent that it affects audiovisual perception. If IDS enhances visual speech cues more than ADS does, infants may more readily respond to mismatching auditory and visual information in IDS than in ADS. Based on previous research (Englund & Behne, 2005; Englund & Behne, 2006), our study focused on infants from birth to 6 months and used an approach similar to the one reported in Dolscheid, Hunnius, Casasanto, and Majid (2014) and close to the one in Kubicek *et al*. (2014) and the second experiment in Shaw, Bart, Depowski, and Bortfeld (2015). Infants faces were video recorded while they gazed at a computer screen, and video recordings were timed manually to measure infant visual fixation.

Based on previous research (e.g., Kuhl & Meltzoff, 1984; Patterson & Werker, 1999; Walton & Bower, 1993), we predicted that infants would look longer at congruent than incongruent audiovisual speech. Similarly, consistent with previous findings (Lewkowicz, 1996; Lewkowicz & Hansen-Tift, 2012), we predicted that infants would look longer at audiovisual IDS than ADS. If visual cues are more prominent in IDS, the difference in fixation to congruent and incongruent IDS syllables should be larger than the same difference in ADS syllables, resulting in an interaction. The current study design included visual fixation time as a dependent variable and two within subjects' factors: "speech type" (ADS and IDS) and "congruence" (congruent and incongruent). An additional between subjects' factor was used with two levels corresponding to two different semi-randomised orders of the stimuli (Order 1 and 2).

Participants

Since our previous study was based on findings for IDS to infants from 0 to 6 months old (Englund & Behne, 2006), twenty infants (10 boys and 10 girls) from 2 to 6 months (mean age 4.9 months) old were included in the study. Mean age was 4.9 months for girls and 5.0 for boys.[1] Infants and mothers from monolingual Norwegian families were recruited through local health care centers and acquaintances. They received no compensation for participation. All were healthy full-term infants with no reported auditory or visual impairment. While one infant was the fifth and four others were the second child, the rest were first born.

Experimental stimuli

IDS and ADS speech stimuli were developed from audiovisual recordings of an adult, native Norwegian, female speaking to her 5-month-old son and to an adult. She was asked to repeat syllables as if she were teaching the infant a new word consisting of one syllable repeated ten times. Likewise, she was asked to repeat syllables as if she were teaching an adult a new word. She was audio-, and video-recorded while completing ten repetitions of each of the CV-syllables /pa:/, /pi:/ and /pu:/ while sitting face-to-face with the listener. Studies of head movement in IDS have evaluated translation and rotation of the head in ADS and IDS and although one study observed more head movement in ADS compared to IDS (Shepard, Spence & Sasson, 2012), it has more widely been found that mothers move their heads more when speaking to their infants than when speaking to an adult (Smith & Strader, 2014). For this reason, the mother was instructed to avoid moving her head when articulating the syllables. Similarly, smiling may influence results, and the mother was asked not to smile throughout recordings. Separate recordings were made of audio and video. A video recording camera (Sony DCR-TRV50E digital video camera recorder with 1550 × 970 XGA resolution and a portable DAT-audio recorder was used (Sony Digital Audio Tape recorder Walkman TCD–D8 with a Shure dynamic headset microphone, model WH20). From the audiovisual recordings, three syllables /pa:/ /pi:/ and /pu:/ from both ADS and IDS were selected. The selection was based on the vowels' similarity to their counterparts from Englund and Behne (2005). This means for ADS the selected/a:/ fell within +/− 20 Hz of 692 Hz for F_1 and 1364 for F_2,/i:/ fell within +/− 20 Hz of 448 for F_1 and 2241 of F_2,/u:/ fell within +/− 20 Hz of 486 for F_1 and 1288 of F_2. Similarly, for IDS the selected/a:/ fell within +/− 20 Hz of 693 for F_1 and 1654 for F_2,/i:/ fell within +/− 20 Hz of 446 for F_1 and 2238 for F_2 while/i:/ fell within +/− 20 Hz of 505 for F_1 and 1466 for F_2. This resulted in six audio syllables. Each of these (e.g., /pa/) was paired with either its corresponding video-syllable (/pa/) to make a congruent audiovisual syllable, or matched with one of the other video-syllables (/pi/ or /pu/) to create an incongruent audiovisual syllable. In cases where syllables were not aligned in duration, the longest visual syllable was selected and the auditory syllable was stretched in duration at the end, prolonging the vowel. This was done within each speech type. Mean fundamental frequency f_o based on (Titze, Baken, Bozeman *et al*., 2015)) for the vowels within the stimulus syllables was measured and syllables selected in ADS and IDS fell within +/− 20 Hz of 247 Hz. This resulted in 18 combinations of auditory and visual stimuli. All CV-syllables were selected as full syllables with an initial/p/ so only the vowel varied between /pa:/, /pi:/ and /pu:/. Englund and Behne (2005) demonstrated higher first, second and third formant frequencies (F_1, F_2 and F_3) in IDS than ADS for / a:/ and /u:/, but F_2 in/i:/ to be equal in the two types of speech. As such the current experiment contributes to study IDS which is collected from speech from a mother to her infant, where IDS samples are chosen from this speech in a way so as to isolate particular features of IDS resembling those found in previous studies of Norwegian IDS.

Procedure

The experiment was carried out in participants' homes, typically in the living room. Background noise was reduced or eliminated (e.g., TV turned off, windows closed, dish washer and other electronic equipment turned off) and after signing a consent form, the mother sat on a chair or a sofa, with the infant on her lap. A portable computer (Compac, PP2140 with screen resolution 1024 × 768 (XGA)) was placed on a table directly in front of the infant. To obtain a seating position where the infant's nose was in horizontal line to the center of the screen, cushions were used; one to sit on and/or one as a backrest. As consequence, in almost all cases (n = 18) direct bodily contact between mother and infant was limited. If daylight from a window fell on the screen distorting the resolution, seating was altered so as to achieve maximum possible resolution. The screen was approximately one-half meter from the infant's face, resulting in a perceived image of the face on the screen approximately 9 × 10 cm. A digital video camera (Sony, DCR-TRV50E) was mounted directly above the computer screen and adjusted to record the infant's face during the experimental session.

Mothers were told that their infants would be watching a video with congruent and incongruent auditory and visual information, and that the infant's visual fixation would simultaneously be video recorded. Blinding the parent may be more puzzling for an infant in a naturalistic setting compared to a laboratory. At home an infant is accustomed to the environment and the introduction of anything new can more readily be given focus, like headphones and wires. For most, visiting a laboratory is a new situation, and the headphones may be overlooked as one of many new stimuli requiring attention. Experience based on running a pilot supported this conclusion and lead us to instruct the mother to watch the video together with the infant, but to sit quietly and avoid any responses to what was going on at the screen. Further, she was instructed to smile reassuringly and look back at the screen should the infant become fussy. A subjective evaluation of the video recordings confirmed that the mothers followed this instruction.

Each trial began with a colorful underwater screensaver paired with instrumental music presented for 10 s to catch the infant's gaze. This was followed by 30 presentations of the same audiovisual syllable, each lasting one second. Together with the introductory screensaver, one trial lasted 40 s. This was done for all syllables. Each trial was presented only once to each infant. Two semi-randomised versions were made of the stimuli, one with IDS as the first trial and the other with ADS as first trial. One-half of the infants was presented with Order 1; the other half was presented with Order 2. The total duration of the experiment was 12 m.

Infants' faces were video recorded while watching the stimuli. Two adult trained raters who were naive to the purpose, conditions and stimuli of the study evaluated video recordings while logging the time infants were fixating on the screen with a timer. In order to avoid losing focus on the infant's face while the camera was running, one could not zoom in too close, resulting in the size of the infant face available to the raters from the recordings being approximately 6 × 7 cm. For each infant within a trial, time was measured from when the infant's gaze was fixated to the middle of the screen with eyes open at any point during a trial until the infant looked away at any point during the trial. There were slight differences between infants with respect to physical activity. Although limited, there were differences in the angle of the neck and the angle from which the infant was gazing at the screen, approximately 10–15 % degrees variation. All fixation within a trial was added to fixation time. If an infant looked away from the screen and then refocused at any time during a trial, fixation time both before and after refixation were added to the sum of fixation for that particular trial. If an infant did not look at the initial screen saver, but looked at the screen during a trial, all fixation time throughout the trial was recorded. If an infant looked away from the screen for two consecutive trials, the procedure would be discontinued for that participant. The two raters first rated all trials and a random selection of 30% was rated by a third rater. In the rare cases where raters' judgements differed, mean fixation time was calculated. To more thoroughly justify that IDS stimuli contained visually enhanced cues, an adult female rater with no knowledge of the purpose of the study evaluated which syllable could be perceived in each trial from the visual stimuli only.

RESULTS

With all infants completing the procedure, there was no attrition in the current study. In the current methodological approach, refixations were included. To get an impression of infants' engagement in the task, relative fixation time defined by percentage fixation of the maximum 30 s was calculated and is presented in Fig. 1 and Fig. 2 for Order 1 and 2 of the stimuli, respectively.

The exact combinations of audio and visual syllables as well as the mean fixation times to the two versions are presented in Table 1.

To test the possible difference between the raters, a Cronbach alpha test was used, demonstrating a high inter-rater reliability of 0.997. Mean fixation times and variance for congruent and incongruent trials in IDS and ADS are presented in Fig. 3.

A general linear model with mixed repeated measures analysis was carried out with two within subjects' factors (speech type and congruence) and one between subjects' factor (order) and infants' mean fixation time as dependent variable. As illustrated in Fig. 3, the analysis revealed a reliably longer fixation time in IDS (μ = 17.8 s, *SD* = 6.4 s) than ADS (μ = 15.8 s, *SD* = 6.2 s) [F (1.19) = 16.52, $p < 0.001$], and a reliably longer fixation time in incongruent trials (μ = 18.2 s, *SD* = 6.0 s) than in congruent trials (μ = 15.4 s, *SD* = 6.5 s) [F (1.19) = 24.37, $p < 0.001$]. Although the mean difference in fixation time between congruent and incongruent trials in IDS ((IDS congruent trials (μ = 15.9 s, *SD* = 6.6 s) and IDS incongruent trials (μ = 19.6 s, *SD* = 5.8 s)) was double that in ADS ((ADS congruent trials (μ = 14.8 s, *SD* = 6.4 s) and ADS incongruent trials (μ = 16.8 s, SD = 6.0 s)), no interaction was observed between speech type and congruence [F (1,19) = 1.86, $p = 0.089$]. In addition, the between subjects' factor (order) had no reliable effect.

Infants looked longer at IDS than ADS and longer at incongruent than congruent audiovisual syllables in both IDS and ADS. Power of the factors within the repeated measures analyses was tested, revealing speech type and congruence to have a high observed power of 097 and 0.99, respectively. The interaction effect had an observed power of 0.25. Results from the rating of visual stimuli demonstrated that syllables were correctly detected in 79% of the IDS trials and 91% of the ADS trials.

DISCUSSION

The current study was designed to investigate the impact of visual enhancement in IDS on audiovisual mismatch detection in a naturalistic setting. Our expectation of longer fixation time for IDS compared to ADS was confirmed. Likewise, we expected infants to look longer at congruent compared to incongruent syllables. In addition, we expected the magnitude of the difference between congruent and incongruent syllables to be larger in IDS compared to ADS. Unexpectedly, infants looked longer at audiovisually incongruent syllables, and although fixation time to incongruent syllables seemed to be a bit higher in IDS than ADS this was not a reliable interaction.

Fixation times and audiovisual incongruence

By displaying the opposite, our results challenge findings that infants look longer at congruent audiovisual syllables (Aldridge *et al.*, 1999; Kuhl & Meltzoff, 1982; Kuhl & Meltzoff, 1984; Patterson & Werker, 1999, 2003; Walton & Bower, 1993). Traditionally, these studies use a paradigm with two screens, one presenting the matching and one presenting the non-matching visual stimulus together with a speech sound (Kuhl & Meltzoff, 1982, 1982; Patterson & Werker, 1999, 2003). The current task leaves infants with no option, presenting each stimulus repeatedly

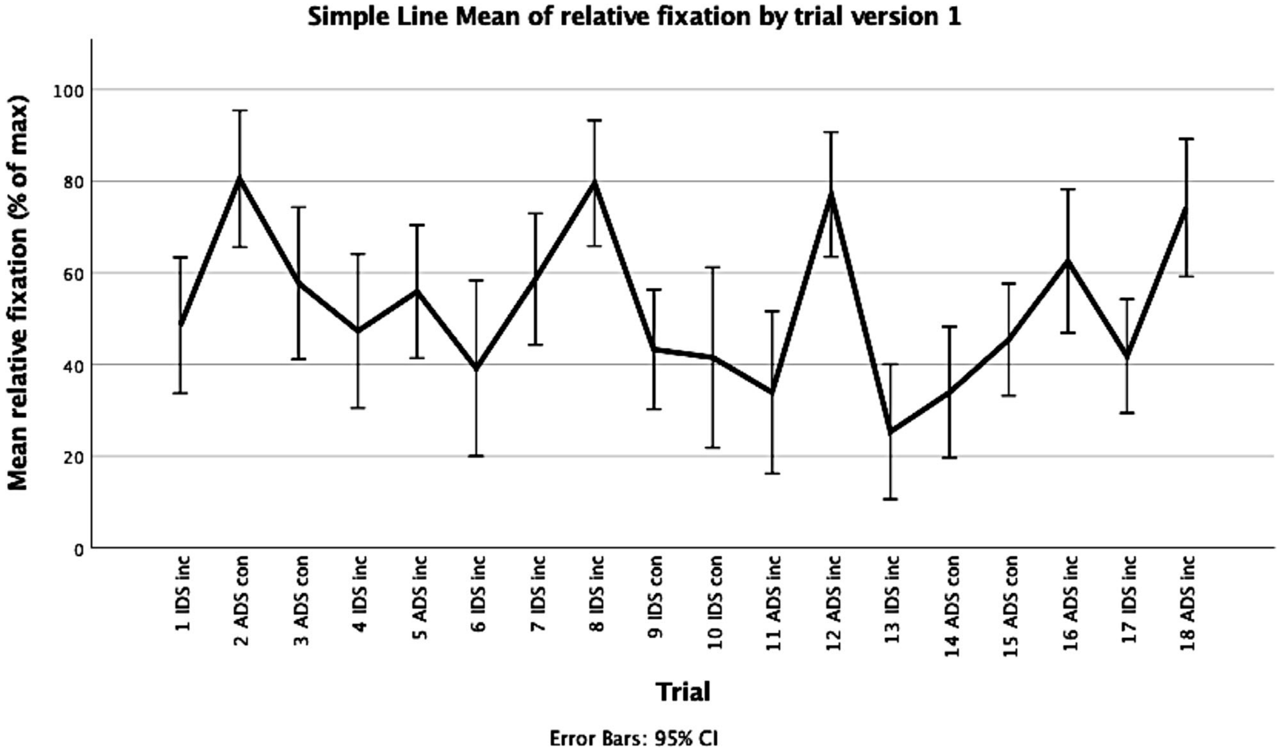

Fig. 1. Percentage of maximum possible fixation time for all 18 trials for order 1 of the stimuli, including error bars at 95% CI. On the x-axis, the content of each trial is indicated, by IDS/ADS, and con/inc = congruent/incongruent.

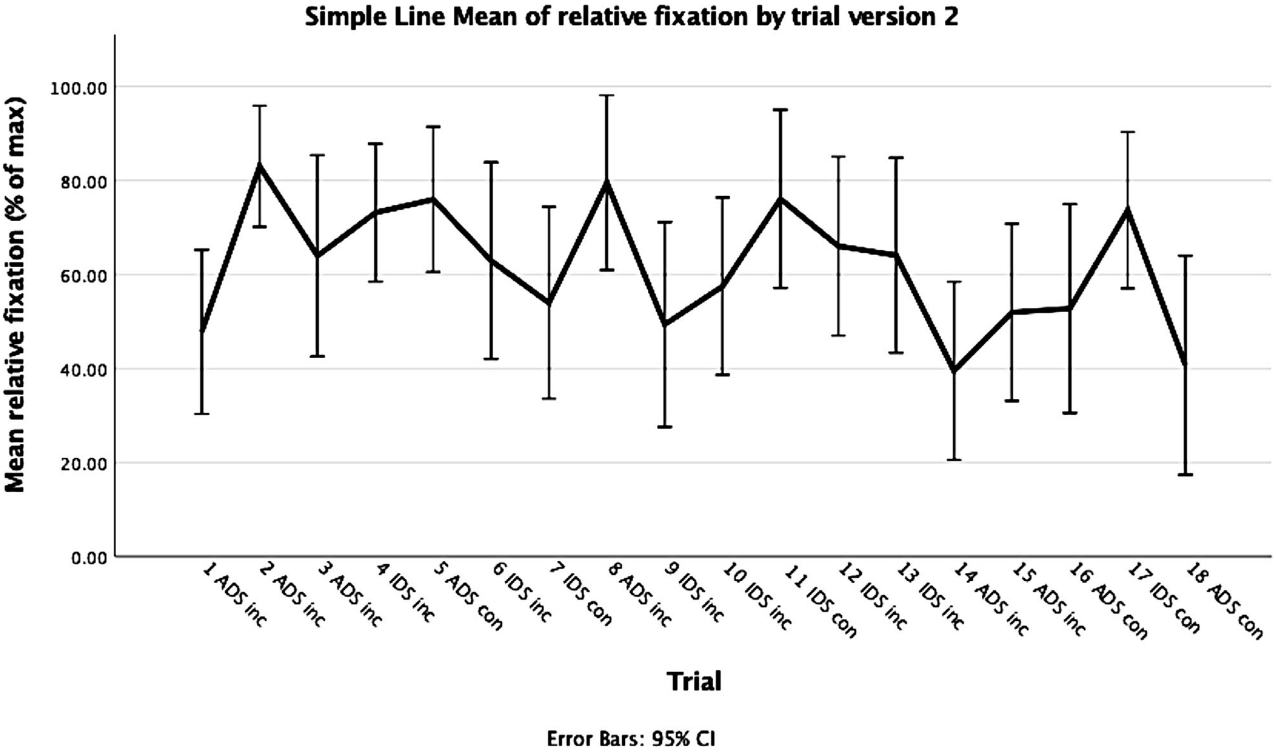

Fig. 2. Percentage of maximum possible fixation time for all 18 trials for order 2 of the stimuli, including error bars at 95% CI. On the x-axis, the content of each trial is indicated, by IDS/ADS, and con/inc = congruent/incongruent.

Table 1. *Audiovisual combinations and fixation times for the stimuli*

	Version 1		Version 2	
Trial	Audio-Visual	Mean fixation	Audio-Visual	Mean fixation
1	/pa:-pi:/	13.98	/pi:-pa:/	15.9
2	/pu:-pu:/	23.81	/pa:-pi:/	25.8
3	/pa:-pu:/	17.36	/pu:-pi:/	19.8
4	/pu:-pi:/	14.63	/pa:-pi:/	22.5
5	/pu:-pa:/	16.33	/pa:-pa:/	23.2
6	/pa:-pu:/	11.57	/pu:-pi:/	19.2
7	/pu:-pa:/	17.83	/pi:-pi:/	16.6
8	/pi:-pi:/	24.05	/pa:-pu:/	24.2
9	/pa:-pa:/	12.12	/pi:-pu:/	15.6
10	/pu:-pu:/	11.73	/pa:-pu:/	17.5
11	/pi:-pu:/	10.47	/pu:-pu:/	23.1
12	/pu:-pi:/	23.71	/pi:-pa:/	20.4
13	/pi:-pa:/	7.44	/pu:-pa:/	19.5
14	/pi:-pi:/	10.56	/pu:-pa:/	13.1
15	/pa:-pa:/	13.97	/pi:-pu:/	15.9
16	/pa:-pi:/	19.32	/pu:-pu:/	16.1
17	/pi:-pu:/	12.54	/pa:-pa:/	22.5
18	/pi:-p:/	22.28	/pi:-pi:/	13.5

Notes: Combinations of audio and visual syllables and mean fixation time in seconds for the two versions of the experimental stimuli in the order they appeared. Audio syllable shown first, and visual syllable presented second in a pair.

on one screen, making a direct comparison with previous studies difficult. Nonetheless, although an infant cannot choose between two screens, there is still the choice between looking at the stimuli or looking away. In that sense, fixation time still reflects a choice. Even so, we can raise the question of what motivates the infant to look at the screen (Houston-Price & Nakkai, 2004). Does it simply reflect the infant's curiosity? Infrequent in natural speech interaction, a mismatching audiovisual syllable can be said to represent a novel situation. In the same vein, it is plausible that the two tasks elicit different processes. Two screens present a more complex situation than one. This complexity may lead the infant to focus on the familiar, that is, what is easier to process. To the extent that infants have experience with processing it, congruent information may be easier, and consequently does not require extra attention and fixation time will be low. Incongruent information, on the other hand, may require extra perceptual capacity and fixation time will be high. On the contrary, being presented with one screen is simpler and may leave perceptual capacity for what is more challenging, namely incongruency. This can account for the longer fixation times to incongruent stimuli in the current results. In addition, there is the possibility that results that seem to be different and that come from two different experimental paradigms may indicate the same perceptual phenomenon. In the word learning literature a switch task presents infants with familiarised audiovisual word-object pairings on one screen (Stager & Werker, 1997), and learning is indicated by longer looking at pairs where word and object mismatch. On the other hand, in a preferential two-screen task looking longer at the screen with the correct or matching object to that of a spoken word also indicates learning (Halberda, 2003). So, both looking longer at a match and looking longer at a mismatch indicate learning. In this vein, looking longer at audiovisual congruence in a two-screen paradigm and looking longer at audiovisual incongruence in a one-screen paradigm may both reflect that infants can differentiate between congruence and incongruence.

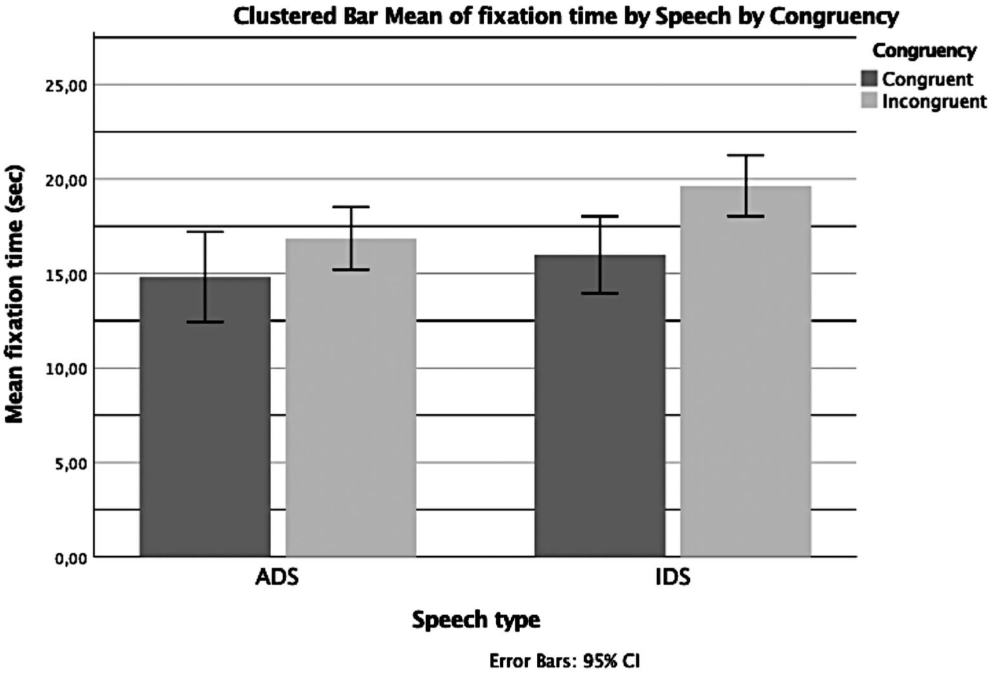

Fig. 3. Duration of fixation time in IDS and ADS for congruent and incongruent audiovisual cues. Mean duration for the two speech types, including error bars at 95% confidence interval.

Repetition of similar stimuli can lead to a learning effect with decreased reaction times over trials, (e.g., Warren & Morton, 1982). With an unequal number of trials for two factors, any difference between them can potentially be the result of this learning effect. Here, we used a higher proportion of incongruent compared to congruent stimuli. Many incongruent trials would lead to expectations for more incongruent trials and the learning effect should have reduced fixation time for this stimulus category. When encountering an unexpected congruent syllable, the result would be longer fixation time for congruent syllables. Since the opposite was observed, this explanation is unlikely.

A related question is how the two orders of the stimuli may have affected perception of congruence. Order 1 had one incongruent and then two congruent syllables, while Order 2 had three incongruent syllables in a row. Despite this, judging from Figs. 1 and 2, as well as evaluation of Table 1, a comparable pattern of fixation is evident during these initial trials. Furthermore, analyses showed order to be a non-significant factor.

Recent research point out that in cases with repeated presentation of incongruent stimuli, unbinding of auditory and visual cues may happen (Nahorna *et al.*, 2012; Nahorna, Berthommier & Schwartz, 2015). In this vein, the generally longer fixation times for incongruent stimuli is better explained as the result of an inability to perceive the auditory and visual information as one.

Fixation times and IDS

As expected, infants looked longer at audiovisual IDS than ADS. This has previously been related to intonational patterns with a wide f_o range and a generally higher pitch (Fernald & Kuhl, 1987). In the same vein, duration is a preferred cue in perception (Bohn & Polka, 2001) and vowels are generally longer in IDS compared to ADS (Bernstein Ratner, 1984; Englund & Behne, 2005). As vowel duration, f_o and f_o range were similar between speech types, this explanation cannot be supported.

It has previously been observed that when hearing a speech sound, kinematic information presented by head movements lead infants to look longer at a visual match, but the same information presented by face movements leads them to look longer at a visual mismatch (Kitamura *et al.*, 2014). Head movements were controlled in the current experiment, and although studying prosody, our results are somewhat in accordance with the previous study, revealing longer fixation at a mismatch based on face movements. The current findings add to this knowledge with results from younger infants, shorter speech stimuli as well as a naturalistic setting.

One of the possible kinds of information provided by visual cues in IDS is emotional information provided by smiling. The selective raising of formant frequencies in the current stimuli could have been the result of smiling. Indeed, research has pointed out that mothers use three distinct facial expressions with emotional intent while speaking to their infants, two of which include smiling (Chong *et al.*, 2003). Positive vocal affect may also have been conveyed by proportion of high frequencies in the spectrum (Banse & Scherer, 1996). When smiling, the mouth widens and lips retract, resulting in a shortened vocal tract and an increase in all formants (Fagel, 2010). This will have contrasting consequences for rounded compared to unrounded vowels. For the vowel /u:/, a smile results in significantly higher F_3, while for /a:/ and /i:/ it did not. So, lip protrusion decrease more from a smile if the vowel is inherently more protruded (Fagel, 2010). A recent study has confirmed no difference between Norwegian IDS and ADS in F_3 for the unrounded /e:, ɛ/ (Englund, 2018). However, the speaker producing materials for the current experiment was instructed not to smile and judging from visual inspection, she did not. Further, research has shown that one may attach more weight to visual input from a smiling than austere speaker (Traunmüller & Öhrström, 2007). If she had been smiling, as a result, visual articulatory cues could have been more prominent. Note that our attempts to control smiling in the current stimuli do not rule out the possibility that the speaker's face is more dynamic in general when producing IDS. From the second experiment in Shepard *et al*. (2012), authors conclude that moving faces of adults speaking IDS provide cues about emotion or intentions. The IDS stimuli in their experiment were made from recordings of speakers asked to provide realistic samples of the way adults speak to infants, implying long stretches of speech, with ample time for perceivable face movement. Each of the current syllables were one second long and less emotion or related information may have been available. Consequently, we have to search for alternative accounts of the effect.

Perceptual challenge hypothesis

Above, we mentioned incongruency as a challenge to the perceptual system. The same can be argued for IDS. The nature of speech is such that it varies along a scale of hyper and hypoarticulation. What determines the degree to which it is one or the other is the perceived need for clear speech in order to get the message across to the perceiver. While by many accounts, IDS is characterized as easier for an infant to learn from than ADS by being hyperarticulated (Gallaway & Richards, 1994; Kuhl *et al*., 2008; Liu, Kuhl & Tsao, 2003), the opposite can also be argued. Recent findings suggest that IDS is characterized by a smaller vowel space, and thereby it is not hyperarticulated, but *hypoarticulated* (Englund, 2018; Martin *et al*., 2015), making it inherently more difficult, (requiring more time) to perceive. This counters findings showing that computer models learn phonetic categories better based on auditory IDS than ADS input (de Boer & Kuhl, 2003). However, computer modeling of speech learning seems to work best based on a fixed set of categories to be learned (Armstrong & Antetomaso, 2011). Infants hardly know the number of categories they are about to learn, challenging the use of computer modeling to argue that IDS is easier to learn from. Although not directly relevant for infant perception, the additional rating of the current stimuli showed that IDS stimuli were more demanding to perceive. We observed that an adult rater found it easier to detect syllables from the current ADS than IDS visual stimuli which may either suggest that IDS in the current experiment was non-representative or that the visual clarity of syllables was reduced in IDS. However, as IDS is usually not spoken to adults, in particular in face-to-face communication, the adult rater may have found the visual IDS unfamiliar compared to ADS and despite the subtlety of visual cues in speech, this may still explain that it was easier for the rater to name the ADS syllables based on visual stimuli.

A complementary way in which IDS may be challenging is by its high variability as shown, for example, in a recent study where IDS vowels were compared to vowels from carefully read speech by the same mothers. Findings for IDS included larger variability and vowels that were acoustically farther apart (Miyazawa, Shinya, Martin, Kikuchi & Mazuka, 2017). Variability is related to hypoarticulation in that as clusters with exemplars of a vowel increase in size, their effective borders overlap, making them less discriminable. If IDS is perceptually challenging by its large variability, and infants still learn phonetic categories from it, it is reasonable to assume that perceptual challenge is not detrimental to learning. In fact, studies of second language acquisition affirm that presenting more variable materials during learning leads to development of more robust categories and that these categories are maintained longer (e.g., Lively, Pisoni, Yamada, Tohkura & Yamada, 1994; Sadakata & McQueen, 2013; Wong, 2014). The current study is not directly comparable since it does not study word learning, studies younger infants, the two speech types which are similar in f_o and f_o range, and articulatory visual cues are included. On the contrary, compared to running speech one can argue that the present stimuli have low levels of variability. Regardless, the current IDS stimuli may be more complex than ADS. In general accounts of learning, low levels of complexity lead to habituation, which in turn may cause low attention and counteract learning (Mather, Schafer & Houston-Price, 2011). Consequently, more complexity counters habituation. In this way, perceptual challenge may promote learning. With the current procedure infants may have become less habituated to incongruent stimuli and IDS stimuli, both of which can be viewed as perceptually challenging.

Strengths and limitations

The stimuli used in the current experiment were not spontaneous, since they are based on recordings from a mother speaking to her infant and selected based on particular criteria to reflect the actual characteristics of IDS spoken to Norwegian infants, with a smaller vowel space than that for ADS.

However, the naturalistic setting provided a high degree of ecological validity in the current study. Infants interact with the environment wherever they are, yet their home environment represents where they interact with family members. Despite this, a naturalistic setting comes at a price, and in the current study it may have affected the results. Although caution was taken to keep physical conditions controlled during data collection, in practice lighting in the room may have differed between participants due to testing being conducted at days with cloudy or sunny weather. Similarly, noise conditions may have been non-identical between participants as a result of different living areas, for example, living close or far away from a main road.

To a large extent, the current experimental setting was naturalistic, and although the visual stimuli did include a small microphone head set, it was very small and not prominent. Any effects of this would be a general effect and could therefore not systematically have affected the independent variables.

A common problem for perception experiments with infants is decay in fixation times caused by habituation and the use of two orders of the stimuli may have remedied some of the possible effects. Judging from Figs. 1 and 2, although distinct for the different trials, infants were engaged throughout the experiment, relatively independent of the order of stimuli.

Parents were not blind to the stimuli and subtle reactions from the mothers may have affected the infants. To what extent this may have been the case is difficult to determine because the mother's potential contingent behaviour is difficult to define, and because adults may have different degrees of compliance with the experimenter's instructions. Note that the last point also constitutes a problem for laboratory studies. The first point is important in that a systematic bias is possible. With direct bodily contact between mother and infant, there is the possibility that the mother's subtle movement may affect the infant. To ensure a direct angle to the computer screen, cushions were used as a seat and backrest for the infant and therefore, for almost all dyads, there was no direct contact between mother and infant. In addition, the mother was instructed to minimize movement and only respond if infants became fussy. Therefore, the mothers' influence on the infants' responses cannot be sole explanation for our findings.

An additional strength of the current study is the simple paradigm used which minimises attrition. In the same vein, the use of relatively inexpensive equipment makes more global replications possible.

CONCLUDING REMARKS

By using audiovisual stimuli with a simple task in an ecologically valid setting, the current study shows that infant gaze is captured more by IDS compared to ADS and more by incongruent compared to congruent stimuli. Although a larger sample size might have led to a significant interaction, incongruent syllables were perceived comparably in both speech types. Stimuli characteristics known to be responsible for infants' preference of IDS were controlled for in the current experiment, hence the results are interpreted as demonstrating that IDS and incongruent stimuli capture gaze, not due to their simplicity but perhaps by posing a challenge to infants' perception.

NOTE

[1] Thanks to Helga Valør Wist, Linda Rydningen and Tine Stokke who recruited infants and collected the data and to Helga Valør Wist for preparing the stimuli. Stimuli were edited by the Norwegian broadcasting company; NRK, Tyholt, Trondheim, Norway.

REFERENCES

Aldridge, M. A., Braga, E. S., Walton, G. E. & Bower, T. G. R. (1999). The intermodal representation of speech in newborns. *Developmental Science*, *2*, 42–46.

Armstrong, T. & Antetomaso, S. (2011). Unsupervised discovery of phoneme boundaries in multi-speaker continuous speech. *Proceedings from: 2011 IEEE International Conference on Development and Learning (ICDL)*, Institute of Electrical and Electronics Engineers, New York.

Banse, R. & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*, 614–636.

Benders, T. (2013). Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior & Development*, *36*, 847–862.

Bernstein Ratner, N. (1984). Patterns of vowel modification in mother-child speech. *Journal of Child Language*, *11*, 557–578.

Bernstein Ratner, N. & Luberoff, A. (1984). Cues to post-vocalic voicing in mother-child speech. *Journal of Phonetics*, *12*, 285–289.

Bohn, O.-S. & Polka, L. (2001). Target spectral, dynamic spectral, and duration cues in infant perception of German vowels. *The Journal of the Acoustical Society of America*, *110*, 504–515.

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

Burnham, D., Kitamura, C. & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*, 1435–1435.

Chong, S. C. F., Werker, J. F., Russell, J. A. & Carroll, J. M. (2003). Three facial expressions mothers direct to their infants. *Infant and Child Development*, *12*, 211–232.

Cristia, A. (2013). Input to language: The phonetics and perception of infant-directed speech. *Language and Linguistics Compass*, *7*, 157–170.

de Boer, B. & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters On-line*, *4*, 129–134.

Desjardins, R. N. & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, *45*, 187–203.

Dolscheid, S., Hunnius, S., Casasanto, D. & Majid, A. (2014). Prelinguistic infants are sensitive to space-pitch associations found across cultures. *Psychological Science*, *1–6*, 1–5. https://doi.org/10.1177/0956797614528521

Englund, K. T. (2018). Hypoarticulation in infant-directed speech. *Applied Psycholinguistics*, *39*, 67–87.

Englund, K. & Behne, D. (2005). Infant directed speech in natural interaction – Norwegian vowel quantity and quality. *Journal of Psycholinguistic Research*, *34*, 259–280.

Englund, K. T. & Behne, D. M. (2006). Changes in infant directed speech in the first six months. *Infant and Child Development*, *15*, 139–160.

Fagel, S. (2010). Effects of smiling on articulation: lips, larynx and acoustics. In A. Esposito, N. Campbell, C. Vogel, A. Hussain & A. Nijholt (Eds.), *Development of multimodal interfaces: active listening and synchrony* (pp. 249–303). Berlin: Springer.

Fernald, A. & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, *10*, 279–293.

C. Gallaway & B. Richards (Eds.) (1994). *Input and interaction in language acquisition*. Cambridge: Cambridge University Press.

Green, J. R., Nip, I. S., Wilson, E. M., Mefferd, A. S. & Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research*, *53*, 1529–1542.

Guellaï, B., Streri, A., Chopin, A., Rider, D. & Kitamura, C. (2016). Newborns' sensitivity to the visual aspects of infant-directed speech: evidence from point-line displays of talking faces. *Journal of Experimental Psychology: Human Perception and Performance. 42*, 1275–1281.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*, B23–B34.

Houston-Price, C. & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*, 341–348.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: The MIT Press.

Kalashnikova, M., Carignan, C. & Burnham, D. (2017). The origins of babytalk: smiling, teaching or social convergence? *Royal Society Open Science*, *4*, 170306. https://doi.org/10.1098/rsos.170306

Kitamura, C., Guellai, B. & Kim, J. (2014). Motherese by eye and ear: Infants perceive visual prosody in point-line displays of talking heads. *PLoS ONE*, *9*, e111467. https://doi.org/10.1371/journal.pone.0111467.

Kubicek, C., Gervain, J., de Boisferon, A. H., Pascalis, O., Lœvenbruck, H. & Schwarzer, G. (2014). The influence of infant-directed speech on 12-month-olds' intersensory perception of fluent speech. *Infant Behavior & Development*, *37*, 644–651.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L. *et al*. Lacerda, F. (1997). Crosslanguage analysis of phonetic units in language addressed to infants. *Science*, *277*, 684–686.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M. & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of The Royal Society B*, *363*, 979–1000.

Kuhl, P. K. & Meltzoff, A. K. (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 & Development;Infant Behavior & Development*, *7*, 361–381.

Lewkowicz, D. J. (1996). Infants' response to the audible and visible properties of the human face. I: Role of lexical-syntactic content, temporal synchrony, gender, and manner of speech. *Developmental Psychology*, *32*, 347–366.

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

Lewkowicz, D. J. & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *109*, 1431–1436.

Liu, H.-M., Kuhl, P. K. & Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, *6*, F1–F10.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y. & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, *96*, 2076–2087.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E. & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, *26*, 341–347.

Mather, E., Schafer, G. & Houston-Price, C. (2011). The impact of novel labels on visual processing during infancy. *British Journal of Developmental Psychology*, *29*, 783–805.

Miyazawa, K., Shinya, T., Martin, A., Kikuchi, H. & Mazuka, R. (2017). Vowels in infant-directed speech: More breathy and more variable, but not clearer. *Cognition*, *166*, 84–93.

Nahorna, O., Berthommier, F. & Schwartz, J. L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustical Society of America*, *132*, 1061–1077.

Nahorna, O., Berthommier, F. & Schwartz, J. L. (2015). Audio-visual speech scene analysis: Characterization of the dynamics of unbinding and rebinding the McGurk effect. *Journal of the Acoustical Society of America*, *137*, 362–377.

Oates, J. (1998). Risk factors for infant attrition and low engagement in experiments and free-play. *Infant Behavior & Development*, *21*, 555–569.

Patterson, M. L. & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior & 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*, 191–196.

Pomerleau, A., Malcuit, G., Chamberland, C., Laurendeau, M.-C. & Lamarre, G. (1992). Methodological problems in operant learning research with human infants. *International Journal of Psychology*, *27*, 417–432.

Sadakata, M. & McQueen, J. M. (2013). High stimulus variability in nonnative speech learning supports formation of abstract categories: Evidence from Japanese geminates. *Journal of the Acoustical Society of America*, *134*, 1324–1335.

Shaw, K., Baart, M., Depowski, N. & Bortfeld, H. (2015). Infants' preference for native audiovisual speech dissociated from congruency preference. *PLoS ONE*, *10*, e0126059. https://doi.org/10.1371/journal.pone.0126059.

Shaw, K. E. & Bortfeld, H. (2015). Sources of confusion in infant audiovisual speech perception research. *Frontiers in Psychology*, *6*, 1844. https://doi.org/10.3389/fpsyg.2015.01844.

Shepard, K. G., Spence, M. J. & Sasson, N. J. (2012). Distinct facial characteristics differentiate communicative intent of infant-directed speech. *Infant and Child Development*, *21*, 555–578.

Smith, N. A. & Strader, H. A. (2014). Infant directed visual prosody. *Interaction Studies*, *15*, 38–54.

Stager, C. L. & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*, 381–382.

Stevenson, M. B., Leavitt, L. A., Roach, M. A., Chapman, R. S. & Miller, J. F. (1986). Mothers' speech to their 1-year-old infants in home and laboratory settings. *Journal of Psycholinguistic Research*, *15*, 451–461.

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

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology*, *5*, 725. https://doi.org/10.3389/fpsyg.2014.00725.

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T. *et al*. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *Journal of the Acoustical Society of America*, *137*, 3005–3007.

Traunmüller, H. & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, *35*, 244–258.

Walton, G. E. & Bower, T. G. (1993). Amodal representations of speech in infants. *Infant Behavior & Development*, *16*, 233–243.

Warren, C. & Morton, J. (1982). The effects of priming on picture recognition. *British Journal of Psychology*, *73*, 117–129.

Weisleder, A. & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*, 2143–2152.

Werker, J. F. & McLeod, P. J. (1989). Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *43*, 230–246.

Wong, J. W. S. (2014). The effects of high and low variability phonetic training on the perception and production of English vowels /e/-/ae/ by Cantonese ESL Learners with High and Low L2 Proficiency Levels. *Proceedings of the 15th Annual Conference of the International Speech Communication Association*, 524–528. Retrieved from https://repository.hkbu.edu.hk/hkbu_staff_publication/ 6234

Zhang, Y., Koerner, T., Miller, S., Grice-Patil, Z., Svec, A., Akbari, D. *et al*. Carney, E. (2011). Neural coding of formant-exaggerated speech in the infant brain. *Developmental Science*, *14*, 566–581.

Received 24 August 2018, Revised 4 July 2019, accepted 20 October 2019